Applications of categorical marginal models in test construction

Kuijpers, R.E.

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

Applications of Categorical Marginal Models in Test Construction

Renske Elisabeth Kuijpers

Tilburg University

Applications of Categorical Marginal Models in Test Construction

ISBN: 978-90-5335-993-8
Printed by: Ridderprint BV, Ridderkerk, The Netherlands.
Cover design: Victoria Schrauwen-Gonzalez

Applications of Categorical Marginal Models in Test Construction

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op

vrijdag 16 januari 2015 om 14.15 uur

door

Renske Elisabeth Kuijpers

geboren op 25 september 1986
te Zevenbergen

Leden van de promotiecommissie:

Promotor:	Prof. dr. K. Sijtsma
Co-promotores:	Dr. L. A. van der Ark
	Dr. M. A. Croon
Overige leden:	Prof. dr. C. A. W. Glas
	Prof. dr. J. A. P. Hagenaars
	Prof. dr. R. R. Meijer
	Prof. dr. M. J. de Rooij

Contents

Chapter 1

Introduction

Measurement is embedded in everyday life; from proud parents measuring their children's height each year, and people regularly recording their body weight, to school teachers grading their students' performance. And what is a good women's magazine without a love or relationship quiz in it? Also, when one wants to obtain a driver's license, both practical driving skills and theoretical traffic knowledge are tested. The theoretical traffic exam is a well-known example of how a construct can be measured by means of a set of items, here multiple-choice questions that ask about traffic rules and require the student to assess and solve typical, practical traffic situations.

Social scientists use tests and questionnaires to measure a variety of different constructs that cannot be observed directly, like depression, anxiety, intelligence, neuroticism, work satisfaction, or attitudes towards religion or abortion. In most standard cases, researchers administering tests to respondents assume that the test-takers do not influence one another's responses, thus they assume that the different respondents' answers are independent. However, answers can also be dependent; for example, the same respondents can be assessed at multiple occasions, respondents can have a personal relation with each other (e.g., mother and daughter or husband and wife), or respondents are members of the same subgroup (e.g., children attending the same school). When observations in a sample are dependent, standard statistical procedures are not sufficient and produce biased results (Bergsma, Croon, & Hagenaars, 2009; p. vi). Methods for analyzing dependent data are available, but many of these methods are based on additional assump-

tions that may not be satisfied in real data so that they can only be applied to a limited number of research questions. A solution is to use marginal models for categorical data (e.g., see Bergsma, 1997; Bergsma et al., 2009; Lang & Agresti, 1994; Molenberghs & Verbeke, 2005). In this dissertation, we focus on marginal modelling for *categorical* data, since most tests and questionnaires that are used in the social sciences use items with discrete item scores.

Categorical marginal models are flexible models for analyzing dependent or clustered categorical data without making specific assumptions about the nature of these dependencies (Bergsma et al., 2009). In this dissertation, categorical marginal models are applied to various research problems in test construction. Standard statistical procedures are often not available, inappropriate to solve the research problem at hand, or are based on restrictive assumptions.

1.1 Categorical Marginal Models

Categorical marginal models handle dependencies in a data set by analyzing entire item-score patterns as a whole rather than analyzing the separate scores on individual items. For example, consider a set of items that measures the degree to which respondents suffer from depression after a major life event like a divorce, the death of a spouse, surviving a life threatening disease, or recovering from an addiction. The scores from, say, 325 respondents on ten items each with three answer categories (e.g., 0 = "disagree", 1 = "neutral", 2 = "agree") can be collected in a ten-dimensional contingency table that consists of $3^{10} = 59,049$ cells. This contingency table has ten one-dimensional marginal tables, each with three cells, showing the frequency distribution of the scores on a particular item. Furthermore, the contingency table has 45 two-dimensional marginal tables, each table having nine cells, showing the joint distribution of the scores on a particular pair of items.

Categorical marginal models impose restrictions on certain marginals of the contingency table in order to test various hypotheses or models. For example, one may be interested in the degree to which items in a test discriminate, and whether an item should be included in a scale. This can be investigated by means of Mokken's (1971) item scalability coefficient H_j, which

Table 1.1: *Univariate and Bivariate Frequencies for Items 1, 2, and 4 of Industrial Malodor Data. Left Column: Observed Frequencies. Middle Column: Expected Frequencies Under Restriction that* $\mathbf{H}_j = \mathbf{.3}$*. Right Column: Expected Frequencies Under Restriction that* $H_1 = H_2 = H_4$*.*

Observed bivariate frequencies

	$X_4 = 0$	
X_1	$X_2 = 0$	$X_2 = 1$
0	250	16
1	172	37

	$X_4 = 1$	
X_1	$X_2 = 0$	$X_2 = 1$
0	16	49
1	30	258

Expected bivariate frequencies for $\mathbf{H}_j = \mathbf{.3}$

	$X_4 = 0$	
X_1	$X_2 = 0$	$X_2 = 1$
0	187.34	38.38
1	154.33	50.93

	$X_4 = 1$	
X_1	$X_2 = 0$	$X_2 = 1$
0	62.31	51.73
1	110.30	172.67

Expected bivariate frequencies for $H_1 = H_2 = H_4$

	$X_4 = 0$	
X_1	$X_2 = 0$	$X_2 = 1$
0	260.39	14.02
1	160.02	38.41

	$X_4 = 1$	
X_1	$X_2 = 0$	$X_2 = 1$
0	15.99	28.53
1	52.19	258.46

Observed univariate frequencies

$X_1 = 0$	$X_1 = 1$
331	497

$X_2 = 0$	$X_2 = 1$
468	360

$X_4 = 0$	$X_4 = 1$
475	353

Expected univariate frequencies for $\mathbf{H}_j = \mathbf{.3}$

$X_1 = 0$	$X_1 = 1$
339.76	488.24

$X_2 = 0$	$X_2 = 1$
514.29	313.71

$X_4 = 0$	$X_4 = 1$
431.00	397.00

Expected univariate frequencies for $H_1 = H_2 = H_4$

$X_1 = 0$	$X_1 = 1$
318.93	509.07

$X_2 = 0$	$X_2 = 1$
488.59	339.41

$X_4 = 0$	$X_4 = 1$
472.84	355.16

Observed scalability coefficients

H_1	H_2	H_4
0.544	0.677	0.674

Expected scalability coefficients for $\mathbf{H}_j = \mathbf{.3}$

H_1	H_2	H_4
0.300	0.300	0.300

Expected scalability coefficients for $H_1 = H_2 = H_4$

H_1	H_2	H_4
0.675	0.675	0.675

should be at least equal to .3 in order to only include well-discriminating items in a scale. In Sijtsma and Molenaar (2002, pp. 82-86), a set of 17 items measuring 828 people's coping strategies regarding industrial malodors (Cavalini, 1992) is used throughout to explain and illustrate the use of Mokken's scalability coefficients. Here, we use part of this data set to illustrate the use of categorical marginal models. According to Cavalini (1992), one of the

scales in the data set consists of the items 1, 2, and 4, measuring the effort to protect the laundry and the inside of the house from the toxic outside air. After dichotomization of the item scores, the corresponding observed H_j coefficients are equal to $H_1 = 0.544$, $H_2 = 0.677$, and $H_4 = 0.674$. As an example, we test the marginal model that for this 3-item scale all three item scalability coefficients H_j are equal to .3 (i.e., $\mathbf{H}_j = \mathbf{.3}$, where $\mathbf{H}_j$ is a vector containing all three H_j's). Restrictions on the marginals are imposed in such a way, that the requirement that all three H_j's are equal to .3 are met. Table 1.1 shows the observed univariate and bivariate frequencies on the left-hand side, and the so-called expected univariate and bivariate frequencies in the middle column. It may be noted that the frequencies in the cells under the marginal model, known as the expected frequencies, are different from the observed frequencies. Using categorical marginal models, the expected frequencies are estimated under the restrictions of the marginal model, such that they are as close as possible to the observed frequencies in the sample, as shown in Table 1.1. Then, the global fit of the marginal model can be assessed; that is, the difference between the observed and expected frequencies is assessed using a likelihood ratio test. The global fit of the marginal model $\mathbf{H}_j = \mathbf{.3}$ equals $G^2 = 210.177$, with $df = 3$ and $p < 0.000$, which indicates that the item scalability coefficients are significantly different from .3. In addition, we tested the marginal model of equal item scalability coefficients (i.e., $H_1 = H_2 = H_4$). On the right-hand side, Table 1.1 shows the expected univariate and bivariate frequencies for the model. The results show that the item scalability coefficients are not equal to each other, since $G^2 = 24.838$, with $df = 2$ and $p < 0.000$.

For estimating categorical marginal models, different estimation methods can be used. We will focus on two methods: the likelihood method and the generalized estimating equations (GEE; Liang & Zeger, 1986) method. The likelihood method, which includes maximum likelihood (ML) estimation (Bergsma, 1997), maximum empirical likelihood (MEL) estimation, and maximum augmented empirical likelihood (MAEL) estimation (Van der Ark, Bergsma, & Croon, 2013; Van der Ark, Croon, & Bergsma, 2011), maximizes the likelihood function under the restrictions of the marginal model. The three likelihood methods ML, MEL, and MAEL, differ in whether or not

they use all possible item-score patterns of a set of items when estimating a marginal model. In contrast to the likelihood method, GEE does not assume a specific probability model for the data. Therefore, GEE is simpler and computationally more straightforward than likelihood estimation. However, GEE has problems with respect to efficiency and accuracy when estimating standard errors of parameters or coefficients (e.g., Agresti, 2013, p. 467; Bergsma et al., 2009, p. vii). In Chapter 5, the two types of estimation methods are compared with respect to different research questions.

Categorical marginal models can be used in a wide range of research situations, for instance, for testing hypotheses involving scalability coefficients in case of dichotomous items (Van der Ark, Croon, & Sijtsma, 2008a), testing marginal homogeneity (e.g., Agresti, 2013, p. 425), assessing the change in marijuana and alcohol use over time among adolescents (Bergsma et al., 2009, pp. 130-148), investigating whether different variables such as age, gender, education, and religiosity have a significant effect on the opinion towards women's lives and roles (Bergsma et al., 2009, pp. 168-171), applying graphical models in research on social mobility (Németh & Rudas, 2013), and investigating the effect of two types of vaccinations on the presence of respiratory problems and headaches in two trial periods (Molenberghs & Verbeke, 2005). Marginal modelling has been applied mainly to testing various content-specific regression models. In the chapters of this dissertation, categorical marginal models are used to solve various psychometric problems in test construction.

1.2 Test Construction

1.2.1 Reliability

The quality of a test can be assessed by means of the test-score reliability. In general, reliability is defined as the degree to which the performance of a respondent on a test stays the same, when the test is administered a second time under exactly identical circumstances. When a test, such as the test measuring depression after a major life event, is perfectly reliable the respondent has exactly the same score when the test is administered a second time under exactly identical circumstances. Hence, no other disturbing

life events happened in between test administrations that influenced the test scores upon repetition, but also the weather outside was the same during both administrations (compare someone's mood when the sun is shining to someone's mood when it is raining cats and dogs), as well as the noise level in the room where testing took place. Since test-score reliability, defined as the correlation between two test replications, cannot be computed on the basis of the data collected in one test administration, it is commonly estimated by means of one of the available methods that approximate reliability. The most frequently used reliability estimation method is Cronbach's coefficient alpha (Cronbach, 1951). Almost every published psychological test reports the reliability by means of this coefficient (Sijtsma, 2009). Most researchers only report the point estimate of coefficient alpha, but do not take the uncertainty of the estimate into account. In Chapter 2, we use the categorical marginal modelling approach to derive three hypothesis tests for Cronbach's alpha, and compare the approach to several alternative methods for testing alpha.

Even though Cronbach's alpha is the most common reliability estimate, few researchers seem to realize it is a lower bound to the reliability (e.g., Lord & Novick, 1968). Better alternatives for estimating reliability are available, like coefficient λ_2 (Guttman, 1945) and the greatest lower bound (GLB; Bentler & Woodward, 1980; Jackson & Agunwamba, 1977; Woodhouse & Jackson, 1977). Many researchers use Cronbach's alpha as a measure for internal consistency, which is commonly defined as the degree to which the items in a test measure one and the same construct (Sijtsma, 2009). However, different researchers argued (Cortina, 1993; Schmitt, 1996; Sijtsma, 2009; Sijtsma & Emons, 2011) that Cronbach's alpha in fact does not indicate whether the items measure the same construct. Given the reasons not to use alpha, why did we still construct hypothesis tests for Cronbach's alpha? The answer is that Cronbach's alpha is the most used reliability estimate. Even though it is a lower bound and not the best method to estimate reliability, it is still better if one uses alpha to also report the uncertainty of the estimate than only to report a point estimate.

1.2.2 Scalability Coefficients

Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002), among other model assessment methods, involves an item selection algorithm that can be used to partition a set of items into one or more scales, with each scale measuring one specific construct. For instance, for the test assessing depression after a major life event it might turn out that the test consists of more than one scale. Not only does the test measure the degree to which you are depressed after something horrible happened in your life, but maybe it also measures the fear that something awful will happen to you again. In addition, the test may measure another mental disorder, such as a negative self image.

Three scalability coefficients are used to determine whether or not items form a scale, and as diagnostics to assess the strength of the scales: (1) item pair scalability coefficient H_{ij}, which expresses the strength of the association between items i and j; (2) item scalability coefficient H_j, which expresses how well item j fits with the other items in a test, and also indicates the extent to which item j discriminates between respondents (Sijtsma & Molenaar, 2002, p. 66); and (3) total-scale scalability coefficient H, which expresses the degree to which respondents can be ordered by means of a set of items (Sijtsma & Molenaar, 2002, pp. 36, 39).

Similar to how Cronbach's alpha in general is reported by applied researchers, scalability coefficients usually are reported without standard errors or other measures of uncertainty as well. However, ignoring standard errors can lead to incorrect inferences about which items to include in a Mokken scale, and about the strength of a scale. Although some researchers were able to derive standard errors for (one of the) scalability coefficients, none were able to derive standard errors for coefficients based on large numbers of items. Furthermore, standard errors were not available for polytomous items, but could only be computed for small sets of dichotomous items. In Chapter 3, we derive standard errors for scalability coefficients by means of categorical marginal models. The method for deriving standard errors is extended to polytomous items and large sets of items. In Chapter 4, we assess the bias of the estimates of the scalability coefficients and their standard errors, and the coverage of the corresponding 95% confidence intervals.

1.3 Outline of the Dissertation

In this dissertation, categorical marginal models are applied to various research problems in test construction. Most researchers only report the point estimates of coefficients, that express quality aspects of the assessed tests. We use categorical marginal modelling to construct hypothesis tests and standard errors, since it is important to take the uncertainty of estimates into account.

In Chapter 2, categorical marginal models are used to construct statistical tests for three hypotheses pertaining to Cronbach's alpha, which is the most widely used reliability coefficient in psychological test construction. The newly developed statistical tests rest on fewer assumptions than existing tests, they are especially suited for discrete item scores, and they can be applied easily to psychological tests containing large numbers of items. In a simulation study, the marginal modelling approach is compared to several of the existing tests.

In Chapter 3, the categorical marginal modelling approach is used for deriving standard errors of scalability coefficients that are used in Mokken scale analysis. In contrast to existing methods, the newly developed method allows the computation of standard errors for scalability coefficients for polytomous items and for large numbers of items. In addition, it is demonstrated by means of two real-data examples that ignoring standard errors of scalability coefficients results in incorrect inferences with respect to the constructed scales.

The estimates and the standard errors of the scalability coefficients are derived assuming that the ordering of the item steps in the sample is identical to the ordering of the item steps in the population. If this assumption is violated, the estimates and the standard errors may be biased. In Chapter 4, by means of two simulation studies the bias of the estimates of these scalability coefficients and the bias of the standard errors is investigated, as well as the coverage of the corresponding 95% confidence intervals.

In Chapter 5, it is explored to what extent the two types of estimation methods, the maximum likelihood method and GEE, are appropriate for investigating different types of research questions that prevail in test construction. It is concluded that the maximum likelihood method can be used for all types of research questions but that the method becomes problematic

for large numbers of items. The GEE method is preferred for conventional regression problems but because the method does not readily provide global goodness-of-fit statistics, it is less useful for the type of hypothesis testing as discussed in Chapter 2.

The dissertation concludes with an epilogue, in which we reflect on the main findings of this dissertation, and discuss advantages and disadvantages of the categorical marginal modelling approach. Furthermore, we discuss implications for future research, and consider other remaining issues.

Chapter 2

Testing Hypotheses Involving Cronbach's Alpha Using Marginal Models

Abstract We discuss the statistical testing of three relevant hypotheses involving Cronbach's alpha: one where alpha equals a particular criterion; a second testing the equality of two alpha coefficients for independent samples; and a third testing the equality of two alpha coefficients for dependent samples. For each of these hypotheses, various statistical tests have been proposed previously. Over the years, new tests have depended on progressively fewer assumptions. We propose a new approach to testing the three hypotheses that relies on even fewer assumptions, is especially suited for discrete item scores, and can be applied easily to tests containing large numbers of items. The new approach uses categorical marginal modelling. We compared the Type I error rate and the power of the marginal modelling approach to several of the available tests in a simulation study using realistic conditions. We found that the marginal modelling approach had the most accurate Type I error rates, whereas the power was similar across the statistical tests.

This chapter has been published as Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013). Testing hypotheses involving Cronbach's alpha using marginal models. *British Journal of Mathematical and Statistical Psychology, 66,* 503-520.

2.1 Introduction

In the social and behavioral sciences, psychometric instruments such as tests, questionnaires, and observation scales are used to measure social and behavioral constructs such as depression, quality of life, and social capital. One of the most important criteria to assess the quality of a measurement instrument is test-score reliability. Test-score reliability cannot be computed directly, and in practice reliability is assessed by means of a coefficient that estimates the reliability. The most frequently used coefficient used to estimate reliability is Cronbach's alpha (Cronbach, 1951), with more than 8,000 citations in Web of Science. We denote the population value by ρ_α, and the sample value by r_α. Three important issues to consider when assessing reliability estimates such as alpha are: (1) whether the absolute value equals a particular criterion; (2) testing the equality of the values for two independent samples; and (3) testing the equality of the values for two dependent samples. Each issue can be formulated as a hypothesis that can be tested statistically.

The first hypothesis posits that Cronbach's alpha is smaller than or equal to a criterion c:

$$H_{01} : \rho_\alpha \leq c. \tag{2.1}$$

Rejecting H_{01} indicates that Cronbach's alpha significantly exceeds the required criterion c. Hypothesis H_{01} is relevant for assessing the criteria proposed by Nunnally (1978, pp. 245-246). He argued that tests that are used to make important decisions about individuals should have a reliability of at least .90 or .95, and tests that are used to make decisions about groups should have a reliability of at least .80. For example, if a researcher finds that $r_\alpha = .81$, then due to sample fluctuation ρ_α may be smaller than the desired .80, and the researcher must test hypothesis H_{01} to demonstrate that $\rho_\alpha > .80$.

The second hypothesis posits that the alpha coefficients for two independent groups, g_1 and g_2, are equal:

$$H_{02} : \rho_{\alpha_{g_1}} = \rho_{\alpha_{g_2}}. \tag{2.2}$$

Hypothesis H_{02} is relevant when the two independent groups have been administered the same test or when they have been administered two different

tests. In test construction, equivalence of alpha across norm groups is an important issue. For example, De Fruyt, De Bolle, McCrae, Terracciano, and Costa (2009) compared the reliability of the scales of the NEO-PI-3 (McCrae, Costa, & Martin, 2005) among 24 different cultures, and reported that for the Openness to Experience scale the reliability was considerably lower in the norm samples from Puerto Rico, Uganda, and Malaysia. For the other scales the alphas were equal. However, these claims were not tested.

The third hypothesis posits that the alpha coefficients for two tests, t_1 and t_2, administered to the same sample are equal:

$$H_{03} : \rho_{\alpha_{t_1}} = \rho_{\alpha_{t_2}}. \tag{2.3}$$

Hypothesis H_{03} may be tested when a single test has been administered twice to the same group at different time points or when two different tests have been administered to the same group. Hypothesis H_{03} is important for comparing the alpha of different subscales within samples, but also for longitudinal research when alpha is assessed over time. For example, Jansen, Essink-Bot, Duvekot, and Van Rhenen (2007) compared the psychometric properties, including test-score reliability estimated by Cronbach's alpha, of three health-related quality of life scales administered to the same sample of women just after childbirth and six weeks after childbirth.

For each of the three hypotheses, different statistical tests have been developed. The earliest tests, based on the work of Feldt (1965), were characterized by rather strong assumptions such as continuous data, multivariate normality, compound symmetry, and homogeneity of variance. Later tests, based on the work of Van Zyl, Neudecker, and Nel (2000), relied on fewer assumptions, resulting in the asymptotic distribution-free (ADF) tests (Maydeu-Olivares, Coffman, Garcia-Forero, & Gallardo-Pujol, 2010; Maydeu-Olivares, Coffman, & Hartmann, 2007). Except for the ADF tests, the assumptions are unrealistic because almost all item scores in psychological tests and questionnaires are discrete, typically having two to five ordered integer values. For some statistical tests, especially those pertaining to H_{01}, robustness studies have been done, but for other tests, especially those pertaining to H_{03}, only a few robustness studies have been conducted. We propose an approach to testing the three hypotheses based on marginal modelling (Bergsma, 1997; Bergsma et al., 2009; Bergsma & Rudas, 2002; Lang & Agresti, 1994; see also

Grizzle, Starmer, & Koch, 1969; Forthofer & Koch, 1973). This approach can be used to test all three hypotheses, and only assumes that the item-score patterns follow a multinomial distribution, which renders the approach suitable for discrete item scores. Moreover, we compared the Type I error rate and the power of several available statistical tests and the marginal modelling approach in a simulation study based on discrete data. In contrast to earlier simulation studies using continuous item scores, we used a data generation model that generated discrete item-score vectors, which fits better with practical data analysis. The marginal modelling approach is rather involved, but can be computed using the R-package *cmm* (Bergsma & Van der Ark, 2013). As of version 0.7, the R documentation file `TestCronbachAlpha.Rd` in this package (type `help(TestCronbachAlpha)`) shows how to perform the analyses in this chapter.

This chapter is organized as follows. First, we briefly discuss the available statistical tests for hypotheses H_{01}, H_{02}, and H_{03} (Equations 2.1, 2.2, and 2.3). Second, we describe the marginal modelling approach. Third, we study the Type I error rate and the power of several available tests and the marginal modelling approach for each of the three hypotheses. Finally, we discuss the strengths and limitations of our approach, and we give recommendations for future research.

2.2 Available Statistical Tests

We use the following notation. Let X_j denote the score on item j (with $j = 1, \ldots, J$) with realization x (with $x = 0, \ldots, k$), and let X_+ be the sum of the J item scores; that is, $X_+ = \sum_{j=1}^{J} X_j$. Let σ_Y^2 denote the variance of variable Y. Then, ρ_α is defined as

$$\rho_\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J}\sigma_{X_j}^2}{\sigma_{X_+}^2}\right). \tag{2.4}$$

To compute the sample value of Cronbach's alpha, let $SS(Y)$ denote the sum of squares for variable Y; that is, $SS(Y) = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$, where N represents the sample size. Then r_α is defined as

$$r_\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J} SS(X_j)}{SS(X_+)}\right). \tag{2.5}$$

2.2.1 Hypothesis 1: Testing a Fixed Value of Alpha

Feldt (1965) derived an approximation to the sampling distribution of Cronbach's alpha under the assumptions of classical test theory (Lord & Novick, 1968, Chapter 3) and four additional assumptions: (a) the subjects are a random sample from the population; (b) the items are a random sample from the population of items; (c) in the population, the subjects' true item scores are continuously and normally distributed; and (d) over the entire subjects-by-items matrix, the measurement errors have homogeneous variance, are normally distributed, and are independent of each other and of the true scores. Using a two-factor analysis of variance (ANOVA) model, Feldt derived a one-tailed statistical test for hypothesis H_{01} (Equation 2.1). Under $\rho_\alpha = c$, the test statistic

$$W_1 = \frac{(1-c)}{(1-r_\alpha)} \tag{2.6}$$

follows an F distribution with $(N-1)$ and $(N-1)(J-1)$ degrees of freedom. Feldt (1965) studied the robustness of the statistical test of hypothesis H_{01} against violations of the assumptions. For samples having approximately normally distributed test scores based on 80 dichotomous items, he found that the Type I error rate was close to the nominal Type I error rate, but that the Type I error rate for fewer items needed to be further investigated. The power was not investigated.

Van Zyl et al. (2000) derived distributions of Cronbach's alpha under the assumptions of compound symmetry and multivariate normality of the item scores. Yuan, Guarnaccia, and Hayslip Jr. (2003) relaxed these assumptions and Maydeu-Olivares et al. (2007) made further computational simplifications. They provided an ADF estimator of the standard error of r_α, denoted $\hat{\phi}$. For exact formulas, we refer to the appendix that Maydeu-Olivares et al. provided. Hypothesis H_{01} can be tested by computing the one-sided $1-\alpha$ confidence interval of ρ_α with lower limit $r_\alpha - z_{[(1-\alpha)]}\hat{\phi}$. If criterion c is not included in the confidence interval, then H_{01} is rejected. Except when item scores were extremely leptokurtic, Maydeu-Olivares et al. (2007) found good coverage of the ADF confidence intervals, even when the item scores were discrete.

2.2.2 Hypothesis 2: Testing Equality of Alphas for Independent Samples

Feldt (1969) extended his approach to testing hypothesis H_{01} (Equation 2.6) to hypothesis H_{02}. He used the same assumptions as for testing H_{01} and, without loss of generality, he assumed that $r_{\alpha_{g_1}} \geq r_{\alpha_{g_2}}$ (cf. Kim & Feldt, 2008). Under H_{02}: $\rho_{\alpha_{g_1}} = \rho_{\alpha_{g_2}}$, the distribution of the test statistic

$$W_2 = \frac{1 - r_{\alpha_{g_2}}}{1 - r_{\alpha_{g_1}}} \tag{2.7}$$

can be approximated by a central F distribution. Feldt (1969) provided straightforward but yet long formulas to compute the degrees of freedom for this F distribution. For reasons of space, we do not repeat these formulas here. Hakstian and Whalen (1976) and Bonett (2003) generalized Feldt's procedure to multiple groups.

Under the assumption that the data followed a multivariate normal distribution, Kim and Feldt (2008) investigated the Type I error rate and the power for two groups (comparing the statistical tests proposed by Feldt, Hakstian and Whalen, and Bonett) and for three groups (comparing the statistical tests proposed by Hakstian and Whalen, and Bonett). They reported an absence of substantial differences among the three statistical tests: The Type I error rate was satisfactory in all conditions, whereas the power fluctuated across conditions and was difficult to predict.

Maydeu-Olivares et al. (2010) extended the ADF method for testing H_{01} to H_{02} within a structural equation modelling (SEM) framework; for a detailed discussion of this method, see Maydeu-Olivares et al. (2010). Using simulation studies, they showed that Type I error rates were quite accurate.

2.2.3 Hypothesis 3: Testing Equality of Alphas for Dependent Samples

To test the alpha coefficients of two dependent samples, Feldt (1980) discussed two useful modifications of his 1969 procedure. First, as proposed by Pitman (1939), Feldt (1980) discussed test statistic W_2 (Equation 2.7). Let $r_{t_1 t_2}$ denote the sample correlation between the total scores on test t_1 and

test t_2. Then the modified test statistic equals

$$W_3 = \frac{(W_2 - 1)(N - 2)^{1/2}}{(4W_2(1 - r_{t_1t_2}^2))^{1/2}}.$$

Under H_{03}, W_3 is approximated by a t distribution with $(N - 2)$ degrees of freedom. Second, using the Δ method (Kendall & Stuart, 1969, pp. 231-232), Feldt (1980) proposed to test hypothesis H_{03} by means of W_2, and to adjust both degrees of freedom of the F distribution to

$$v = \frac{N - 1 - 7r_{t_1t_2}^2}{1 - r_{t_1t_2}^2},$$

where v is rounded to the nearest lower integer. For a more detailed discussion of these two procedures, see Feldt (1980).

Alsawalmeh and Feldt (1994) proposed a more refined adjustment of the degrees of freedom of W_2 (Equation 2.7). The formulas for the adjusted degrees of freedom are straightforward but long. For reasons of space, we do not repeat these formulas here. Alsawalmeh and Feldt found that their adjustment resulted in better Type I errors than the two methods Feldt (1980) proposed, especially for small numbers of items. For H_{03}, robustness studies to investigate power have not been done. Hence, the robustness of the tests remains unknown and valid results cannot be guaranteed.

To test H_{03}, Maydeu-Olivares et al. (2010) slightly modified the ADF procedure for testing H_{02}. Again, a SEM framework was used to specify a model for testing the alphas of two dependent samples. For more details about the procedure, we refer to Maydeu-Olivares et al. (2010). Simulations showed that the Type I error rates were considered to be acceptable, but were slightly less accurate when compared to the Type I error rates found for H_{02}. This result may be due to the small sample size used for testing hypothesis H_{03}.

2.3 The Marginal Modelling Approach

The new approach to testing hypotheses H_{01}, H_{02}, and H_{03} (Equations 2.1, 2.2, and 2.3, respectively) is based on marginal modelling (e.g., Bergsma, 1997; Bergsma et al., 2009; Bergsma & Rudas, 2002; see Van der Ark, Croon,

& Sijtsma, 2008a, and Kuijpers, Van der Ark, & Croon, 2013b, for applications of marginal modelling in the context of psychological scaling). J items, each having $k+1$ ordered scores, produce $L=(k+1)^J$ different item-score patterns. Let $\mathbf{n}$ be an $L\times 1$ vector containing the observed frequencies of the L different item-score patterns. For example, a dichotomously scored test consisting of $J=3$ items (denoted by a, b, and c) has $L=2^J=8$ possible item-score patterns and vector $\mathbf{n}$ equals

$$\mathbf{n}=\begin{pmatrix} n_{abc}^{000}\\ n_{abc}^{001}\\ n_{abc}^{010}\\ n_{abc}^{011}\\ n_{abc}^{100}\\ n_{abc}^{101}\\ n_{abc}^{110}\\ n_{abc}^{111}\end{pmatrix}, \tag{2.8}$$

where the subscripts denote the items and the superscripts the item scores. Throughout this chapter, the response patterns are ordered lexicographically: going from $00\ldots0$ to $kk\ldots k$ with the last digit changing fastest, then the penultimate digit changing fastest, and so on, and the digit in the first column changing slowest. The vector $\mathbf{n}$ in Equation 2.8 is used throughout to illustrate the approach.

Marginal models place constraints on the observed frequencies in $\mathbf{n}$. Then the frequencies of an $L\times 1$ vector $\mathbf{m}$ are estimated such that, given these constraints, the null hypothesis being tested holds. The expected frequencies of the item-score patterns under the constraints of the null hypothesis being tested are thus collected in vector $\mathbf{m}$. Suppose that D constraints on the expected frequencies $\mathbf{m}$ are required to satisfy the null hypothesis. Each constraint is a scalar function, so $g_1(\mathbf{m})=d_1$, $g_2(\mathbf{m})=d_2$, …, $g_D(\mathbf{m})=d_D$, where $d_1,\ldots,d_D$ are constants. The scalar functions can be collected in a vector $\mathbf{g}(\mathbf{m})$, and constants $d_1,\ldots,d_D$ can be collected in a vector $\mathbf{d}$, such that

$$\mathbf{g}(\mathbf{m})=\begin{pmatrix} g_1(\mathbf{m})\\ \vdots\\ g_D(\mathbf{m})\end{pmatrix}=\mathbf{d}. \tag{2.9}$$

The constraints in Equation 2.9 constitute the *marginal model*. Let $\widehat{\mathbf{m}}$ be

an estimator of $\mathbf{m}$. The vector $\mathbf{m}$ is estimated under the assumption that $\mathbf{g}(\widehat{\mathbf{m}}) = \mathbf{d}$. The usual estimation method for vector $\mathbf{m}$ is maximum likelihood (ML). The global fit of the categorical marginal model is assessed by the likelihood ratio statistic $G^2 = 2\mathbf{n}^T \log(\mathbf{n}/\widehat{\mathbf{m}})$. If the constraints in Equation 2.9 are true, G^2 has an asymptotic chi-square distribution with D degrees of freedom.

To use marginal models for testing H_{01}, H_{02}, and H_{03}, the three hypotheses should be written as constraints on the expected frequencies (Equation 2.9). This can be cumbersome, and so the process is explained step by step. The first step is to rewrite ρ_α (Equation 2.4) as a function of the expected cell frequencies $\mathbf{m}$. A single general matrix formula using a recursive exp-log notation is used (Bergsma, 1997; Kritzer, 1977). Let $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$ and $\mathbf{A}_5$ be design matrices. We show that if one defines these design matrices in a convenient way and one uses the recursive exp-log notation, then ρ_α and r_α can be written as a function of the expected cell frequencies $\mathbf{m}$ and the observed cell frequencies $\mathbf{n}$, respectively. The generalized exp-log expressions for ρ_α and r_α are

$$\rho_\alpha = \mathbf{A}_5 \exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})))), \tag{2.10}$$

and

$$r_\alpha = \mathbf{A}_5 \exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n})))). \tag{2.11}$$

In Equations 2.10 and 2.11, the vector-valued functions $\exp(\mathbf{y})$ and $\log(\mathbf{y})$ should be read as the exponential and the natural logarithm, respectively, and these functions are applied to each element of an arbitrary vector $\mathbf{y}$. The exponential and the logarithmic functions are used for element-wise multiplication and division of the vectors.

Let $\mathbf{R}$ be a $J \times L$ matrix that contains all L response patterns. The rows of $\mathbf{R}$ correspond to the J different items. The response patterns in $\mathbf{R}$ are in lexicographic order (cf. vectors $\mathbf{m}$ and $\mathbf{n}$). Let $\mathbf{u}_J^T$ be a $1 \times J$ unit vector, let $\mathbf{s}^T$ be a $1 \times L$ vector that contains the sums of all possible item-score patterns stored in $\mathbf{R}$ (i.e., $\mathbf{s}^T = \mathbf{u}_J^T \mathbf{R}$), let $\mathbf{R}^{(2)}$ be a $J \times L$ matrix that contains the squared elements of $\mathbf{R}$, and let $\mathbf{s}^{(2)T}$ be a $1 \times L$ vector containing the squared elements of $\mathbf{s}^T$. The $[2J + 3] \times L$ design matrix $\mathbf{A}_1$ is a concatenation of five

submatrices; that is,

$$\mathbf{A}_1 = \begin{pmatrix} \mathbf{R} \\ \mathbf{s}^T \\ \mathbf{R}^{(2)} \\ \mathbf{s}^{(2)T} \\ \mathbf{u}_L^T \end{pmatrix}.$$

For the three dichotomously scored items a, b, and c (Equation 2.8), we have that

$$\mathbf{A}_1\mathbf{n} = \left(\begin{array}{cccccccc} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline 0 & 1 & 1 & 2 & 1 & 2 & 2 & 3 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline 0 & 1 & 1 & 4 & 1 & 4 & 4 & 9 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}\right) \begin{pmatrix} n_{abc}^{000} \\ n_{abc}^{001} \\ n_{abc}^{010} \\ n_{abc}^{011} \\ n_{abc}^{100} \\ n_{abc}^{101} \\ n_{abc}^{110} \\ n_{abc}^{111} \end{pmatrix} = \left(\begin{array}{c} \sum X_a \\ \sum X_b \\ \sum X_c \\ \hline \sum X_+ \\ \hline \sum X_a^2 \\ \sum X_b^2 \\ \sum X_c^2 \\ \hline \sum X_+^2 \\ \hline N \end{array}\right). \quad (2.12)$$

As the first three elements of the right-hand side of Equation 2.12 show, $\mathbf{Rn}$ produces a vector containing the sum of the scores on items a, b, and c across respondents. Furthermore, the fourth element of the right-hand side of Equation 2.12, $\sum X_+$, equals the sum over N total scores. The next three elements contain the sum of the squared item scores times the observed frequencies, for the items a, b, and c. The eighth element produces a similar element, with the only difference that here the squared sum scores across the different items are used. Finally, the last element gives the total number of respondents in the sample.

The $2(J+1) \times [2J+3]$ design matrix $\mathbf{A}_2$,

$$\mathbf{A}_2 = \begin{pmatrix} \mathbf{O} & \mathbf{I}_{J+1} & \mathbf{u}_{J+1} \\ 2 \times \mathbf{I}_{J+1} & \mathbf{O} & \mathbf{o}_{J+1} \end{pmatrix},$$

is a concatenation of several submatrices, in which $\mathbf{O}$ is a $(J+1)\times(J+1)$ zero matrix, $\mathbf{I}_{J+1}$ is an identity matrix of order $(J+1)$, and $\mathbf{o}_{J+1}$ is a zero vector of length $(J+1)$. When substituting the right-hand side of Equation 2.12 for $\mathbf{A}_1\mathbf{n}$, for the three dichotomous items a, b, and c, product $\exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{n}))$

produces

$$\exp\left[\left(\begin{array}{cccc:cccc:c} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ \hdashline 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \log\left[\left(\begin{array}{c} \sum X_a \\ \sum X_b \\ \sum X_c \\ \hdashline \sum X_+ \\ \hdashline \sum X_a^2 \\ \sum X_b^2 \\ \sum X_c^2 \\ \hdashline \sum X_+^2 \\ \hdashline N \end{array}\right)\right]\right] = \left(\begin{array}{c} N\sum X_a^2 \\ N\sum X_b^2 \\ N\sum X_c^2 \\ \hdashline N\sum X_+^2 \\ \hdashline (\sum X_a)^2 \\ (\sum X_b)^2 \\ (\sum X_c)^2 \\ (\sum X_+)^2 \end{array}\right). \tag{2.13}$$

The design matrix $\mathbf{A}_3$ has three rows (independent of the number of items) and $2(J+1)$ columns:

$$\mathbf{A}_3 = \left(\begin{array}{cccc} \mathbf{u}_J^T & 0 & -\mathbf{u}_J^T & 0 \\ \mathbf{o}_J^T & 1 & \mathbf{o}_J^T & -1 \\ \mathbf{o}_J^T & 1 & \mathbf{o}_J^T & -1 \end{array}\right).$$

Note that 0, 1, and -1 are scalars. For the three items a, b, and c, substituting the right-hand side of Equation 2.13 for $\exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{n}))$, product $\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{n}))$ equals

$$\left(\begin{array}{cccc:cccc} 1 & 1 & 1 & 0 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{array}\right) \left(\begin{array}{c} N\sum X_a^2 \\ N\sum X_b^2 \\ N\sum X_c^2 \\ \hdashline N\sum X_+^2 \\ \hdashline (\sum X_a)^2 \\ (\sum X_b)^2 \\ (\sum X_c)^2 \\ (\sum X_+)^2 \end{array}\right) = \left(\begin{array}{c} \sum SS(X_j) \\ SS(X_+) \\ SS(X_+) \end{array}\right). \tag{2.14}$$

Note that both the second and third elements of the right-hand side of Equation 2.14 equal $SS(X_+)$. Why this is necessary is made clear in the next paragraph.

The design matrix $\mathbf{A}_4$ does not depend on the number of items, and can immediately be written in the general form

$$\mathbf{A}_4 = \left(\begin{array}{ccc} 0 & 1 & -1 \\ 1 & -1 & 0 \end{array}\right).$$

For the three items, substituting the right-hand side of Equation 2.14 for $\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))$, product $\exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))))$ yields

$$\exp\left[\begin{pmatrix} 0 & 1 & -1 \\ 1 & -1 & 0 \end{pmatrix} \log\begin{pmatrix} \sum SS(X_j) \\ SS(X_+) \\ SS(X_+) \end{pmatrix}\right] = \begin{pmatrix} 1 \\ \frac{\sum SS(X_j)}{SS(X_+)} \end{pmatrix}. \qquad (2.15)$$

Note that the scalar 1 on the right-hand side of Equation 2.15 was obtained by dividing two equal quantities.

The design matrix $\mathbf{A}_5$ is a 1×2 row vector containing the number of items divided by the number of items minus 1, and the negative of that element. The general form of matrix $\mathbf{A}_5$ is

$$\mathbf{A}_5 = \begin{pmatrix} \frac{J}{J-1} & \frac{-J}{J-1} \end{pmatrix}.$$

When substituting the right-hand side of Equation 2.15 for $\exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))))$, product $\mathbf{A}_5 \exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))))$ (i.e., Equation 2.11) equals

$$\begin{pmatrix} \frac{J}{J-1} & \frac{-J}{J-1} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{\sum SS(X_j)}{SS(X_+)} \end{pmatrix} = \frac{J}{J-1}\left(1 - \frac{\sum SS(X_j)}{SS(X_+)}\right), \qquad (2.16)$$

where the right-hand side equals r_α (see Equation 2.5). Hence, this shows that Equation 2.11 yields the sample estimate r_α (Equation 2.5).

Now that it has been shown how the general expression for Cronbach's alpha can be rewritten into the exp-log notation, we demonstrate how the first hypothesis, $H_{01} : \rho_\alpha \leq c$, can be expressed in terms of Equation 2.9. Testing H_{01} requires one constraint (i.e., $D = 1$). Writing ρ_α in the recursive exp-log notation (Equation 2.10) and letting $\mathbf{d}$ be the scalar c, facilitates writing $H_{01} : \rho_\alpha = c$ as

$$H_{01} : \mathbf{A}_5 \exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})))) = c.$$

The fit of this marginal model is evaluated by G^2, with $D = 1$ degree of freedom. In general, G^2 pertains to a two-sided test. However, here H_{01} is a one-sided hypothesis, and the value of G^2 at the 2α level is used. For $\alpha = 0.05$, H_{01} must be rejected if $G^2 > 2.71$ (i.e., $p = .10$) and $r_\alpha > c$.

Expressing H_{02} into Equation 2.9 should be done as follows. Let the design matrix $\mathbf{A}_{q_{g1}}$, with $q = 1, \ldots, 5$, be the particular design matrix constructed for the first independent group. The design matrix $\mathbf{A}_{q_{g2}}$ represents

the same qth design matrix that is constructed for the second independent group. For testing the equality of two alphas, the design matrices $\mathbf{A}_1^*$ to $\mathbf{A}_5^*$ are the direct sum of $\mathbf{A}_{q_{g1}}$ with $\mathbf{A}_{q_{g2}}$. Since for each design matrix $\mathbf{A}_q^*$ the procedure is the same, it can be expressed in a general form

$$\mathbf{A}_q^* = \mathbf{A}_{q_{g1}} \oplus \mathbf{A}_{q_{g2}} = \begin{pmatrix} \mathbf{A}_{q_{g1}} & 0 \\ 0 & \mathbf{A}_{q_{g2}} \end{pmatrix}. \tag{2.17}$$

Let $\mathbf{m}^*$ be a $2L \times 1$ vector that contains the expected frequencies in group 1 and group 2, respectively. The vector $\mathbf{m}^*$ can be expressed as

$$\mathbf{m}^* = \begin{pmatrix} \mathbf{m}_{g_1} \\ \mathbf{m}_{g_2} \end{pmatrix}.$$

The vector $\mathbf{n}^*$, which contains the observed frequencies of group 1 and group 2, respectively, is constructed in a similar way. The recursive exp-log expression for $\rho_{\alpha_{g_1}}$ and $\rho_{\alpha_{g_2}}$ collected together in one expression is now

$$\begin{pmatrix} \rho_{\alpha_{g_1}} \\ \rho_{\alpha_{g_2}} \end{pmatrix} = \mathbf{A}_5^* \exp(\mathbf{A}_4^* \log(\mathbf{A}_3^* \exp(\mathbf{A}_2^* \log(\mathbf{A}_1^* \mathbf{m}^*))))). \tag{2.18}$$

For testing null hypothesis $H_{02} : \rho_{\alpha_{g_1}} = \rho_{\alpha_{g_2}}$, the constraint placed on the expected frequencies is that the ρ_αs have to be equal. Let $\mathbf{A}_6$ be a 1×2 vector $(1 \ -1)$. Then, by premultiplying both sides of Equation 2.18 by $\mathbf{A}_6$, it follows that

$$(\rho_{\alpha_{g1}} - \rho_{\alpha_{g2}}) = \mathbf{A}_6(\mathbf{A}_5^* \exp(\mathbf{A}_4^* \log(\mathbf{A}_3^* \exp(\mathbf{A}_2^* \log(\mathbf{A}_1^* \mathbf{m}^*)))))). \tag{2.19}$$

Hypothesis $H_{02} : \rho_{\alpha_{g_1}} = \rho_{\alpha_{g_2}}$ is equivalent to $H_{02} : \rho_{\alpha_{g_1}} - \rho_{\alpha_{g_2}} = 0$. It follows from Equation 2.19 that the marginal model restrictions for H_{02} are

$$H_{02} : \mathbf{A}_6(\mathbf{A}_5^* \exp(\mathbf{A}_4^* \log(\mathbf{A}_3^* \exp(\mathbf{A}_2^* \log(\mathbf{A}_1^* \mathbf{m}^*)))))) = 0. \tag{2.20}$$

To evaluate the fit of the marginal model, G^2 is used with $D = 1$ degree of freedom. Since H_{02} is a two-sided hypothesis, it must be rejected if $G^2 > 3.84$ (i.e., $\alpha = .05$).

To test hypothesis $H_{03} : \rho_{\alpha_{t_1}} = \rho_{\alpha_{t_2}}$, the marginal model as derived for H_{02} has to be adjusted slightly. Stored in a single item-score vector, $\mathbf{n}^\dagger$ contains the frequencies of the item-score patterns of both test t_1 and test

t_2, and $\mathbf{m}^\dagger$ contains the corresponding expected frequencies. For example, if both tests consist of two dichotomous items, then

$$\mathbf{n}^\dagger = \begin{pmatrix} n_{00\ 00} \\ n_{00\ 01} \\ n_{00\ 10} \\ n_{00\ 11} \\ \vdots \\ n_{11\ 10} \\ n_{11\ 11} \end{pmatrix}. \tag{2.21}$$

The vector $\mathbf{n}^\dagger$ is multiplied by $\mathbf{A}_0$, which is a *marginal matrix* (Bergsma, et al., 2009, pp. 52-56). Multiplication with matrix $\mathbf{A}_0$ yields the marginal frequencies of the item-score patterns for both sets of items separately. Let L_1 and L_2 be the number of possible item-score patterns for test t_1 and test t_2, respectively. Let $\otimes$ denote the Kronecker product. The general form of the $(L_1 + L_2) \times (L_1 L_2)$ matrix $\mathbf{A}_0$ is

$$\mathbf{A}_0 = \begin{pmatrix} \mathbf{I}_{L_1} \otimes \mathbf{u}_{L_2}^T \\ \mathbf{u}_{L_1}^T \otimes \mathbf{I}_{L_2} \end{pmatrix}. \tag{2.22}$$

For the example where the two tests contain two items (Equation 2.21), $\mathbf{A}_0\mathbf{n}^\dagger$ equals

$$\left(\begin{array}{cccc:cccc:cccc:cccc} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \hdashline 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right) \begin{pmatrix} n_{00\ 00} \\ n_{00\ 01} \\ n_{00\ 10} \\ n_{00\ 11} \\ \vdots \\ n_{11\ 10} \\ n_{11\ 11} \end{pmatrix} = \begin{pmatrix} n_{00\ ++} \\ n_{01\ ++} \\ n_{10\ ++} \\ n_{11\ ++} \\ n_{++\ 00} \\ n_{++\ 01} \\ n_{++\ 10} \\ n_{++\ 11} \end{pmatrix}$$

After premultiplying vector $\mathbf{n}^\dagger$ by $\mathbf{A}_0$, the two alpha coefficients for the two sets of items are computed using the design matrices in Equation 2.17. Then, when using the marginal model from Equation 2.20, matrix $\mathbf{A}_0$, and $\mathbf{m}^\dagger$, the marginal model for testing hypothesis H_{03} is

$$H_{03} : \mathbf{A}_6(\mathbf{A}_5^* \exp(\mathbf{A}_4^* \log(\mathbf{A}_3^* \exp(\mathbf{A}_2^* \log(\mathbf{A}_1^* \mathbf{A}_0 \mathbf{m}^\dagger))))) = 0. \tag{2.23}$$

G^2 is used to assess the fit of the marginal model with $D = 1$ degree of freedom. Since H_{03} is a two-sided hypothesis, it must be rejected if $G^2 > 3.84$ (i.e., $\alpha = .05$).

2.4 Simulation Study

We compared the Type I error rate and the power of several available statistical tests and the marginal modelling approach under conditions that are relevant in practical test construction. The most important of these conditions is that the simulated item scores are discrete. We expect that under these conditions, the marginal modelling approach and the ADF method, which are based on weaker assumptions, have better Type I error rates than the other statistical tests. However, we expect that the ADF method performs less well in case of small sample sizes, as earlier simulation studies have shown (Maydeu-Olivares et al., 2007, 2010). If the tested hypothesis was in agreement with the chosen population model, the Type I error rate was estimated. If the null hypothesis was not in agreement with the population model, the power was estimated.

2.4.1 Method

The simulation study was set up as follows. We used an experimental design with six independent factors. First, for each cell in the design we constructed a population model for discrete item responses. These population models have the property that the Cronbach's alpha(s) in the population (i.e., ρ_{α}, $\rho_{\alpha_{g_1}}$ and $\rho_{\alpha_{g_2}}$, or $\rho_{\alpha_{t_1}}$ and $\rho_{\alpha_{t_2}}$) can be fixed to a certain required value. Hence, these population models allow the sampling of discrete item scores under the null hypothesis of interest. In psychological testing, most item scores are discrete. Using discrete data rather than continuous data in simulation studies fits better with practical data analysis.

We used a two-step procedure to obtain a population model. In step 1, we used an item response theory (IRT) model to generate item-score vectors. We used the two-parameter logistic model (Birnbaum, 1968) for dichotomous items, and the graded response model (Samejima, 1969) for polytomous items. The location parameters and the discrimination parameters were chosen such that the resulting alpha values were close to the required values. For most cells in the design we generated 200,000 item-score vectors from the IRT model. For design cells that pertain to testing H_{03} for five polytomous items or ten dichotomous items we generated 2,000 item-score

vectors and 160,000 item-score vectors, respectively. The observed frequencies of the sampled item-score vectors were gathered in vector $\mathbf{n}$. In step 2, we used a marginal model to estimate expected item-score vectors under the null hypothesis of interest. The type of marginal model depended on the hypothesis being tested (H_{01}, H_{02}, or H_{03}), the required population values of the alpha coefficient in the design cell, the number of item scores k, and the number of items J. The expected item-score vectors, gathered in $\widehat{\mathbf{m}}$, constituted the population model. Because the IRT model in step 1 already yielded population values of alpha close to the desired value, $\mathbf{n}$ and $\widehat{\mathbf{m}}$ were rather similar.

Next, for each cell in the design, 1,000 data sets were drawn from the population model, so the frequencies $\widehat{\mathbf{m}}$ were used as probability weights. The effects of the following factors on the Type I error rate and the power of the two different approaches were studied:

Statistical Tests. For testing hypothesis H_{01}, we compared Feldt's (1965) method, ADF confidence intervals, and the marginal modelling approach. For testing H_{02}, we compared Feldt's (1969) method, the ADF method, and the marginal modelling approach. For testing H_{03}, we compared the two varieties of Feldt's (1980) method, Alsawalmeh and Feldt's (1994) method, the ADF method, and the marginal modelling approach.

Cronbach's alpha. For studying the Type I error rate, we considered the following conditions: low reliability ($\rho_\alpha = 0.70$), standard-level reliability ($\rho_\alpha = 0.80$), high reliability ($\rho_\alpha = 0.90$), and very high reliability ($\rho_\alpha = 0.95$). Note that for hypothesis H_{01}, $c = \rho_\alpha$; for hypothesis H_{02}, $\rho_{\alpha_{g_1}} = \rho_{\alpha_{g_2}} = \rho_\alpha$; and for hypothesis H_{03}, $\rho_{\alpha_{t_1}} = \rho_{\alpha_{t_2}} = \rho_\alpha$. For studying the power, we considered the following conditions: a standard effect ($\rho_{\alpha_1} = 0.80$, $\rho_{\alpha_2} = 0.70$), a small effect ($\rho_{\alpha_1} = 0.81$, $\rho_{\alpha_2} = 0.80$), and high reliability ($\rho_{\alpha_1} = 0.90$, $\rho_{\alpha_2} = 0.80$). Note that for hypothesis H_{01}, $\rho_\alpha = \rho_{\alpha_1}$ and $c = \rho_{\alpha_2}$; for hypothesis H_{02}, $\rho_{\alpha_{g_1}} = \rho_{\alpha_1}$ and $\rho_{\alpha_{g_2}} = \rho_{\alpha_2}$; and for hypothesis H_{03}, $\rho_{\alpha_{t_1}} = \rho_{\alpha_1}$ and $\rho_{\alpha_{t_2}} = \rho_{\alpha_2}$.

Number of item scores (k). The item scores were dichotomous ($k = 1$) or polytomous ($k = 4$).

Number of items (J). The number of items was $J = 5$ or $J = 10$.

Sample size (N). The sample size was equal to 100, 200, 500 or 1000.

Nominal Type I error rate (α). The nominal Type I error rate was $\alpha = .05$ or $\alpha = .01$.

Instead of varying all factors simultaneously, a standard condition was defined to keep the design of the simulation study manageable. The standard condition was defined as evaluating the Type I error rate for the standard-level reliability and the power for the standard effect, for all statistical tests, for $k = 1$, $J = 5$, $N = 200$, and $\alpha = .05$. The standard case was compared to special cases, and for each special case one of the factors was varied.

The dependent variables were the Type I error rate and the power. Type I error values found in a simulation study are never exactly equal to the nominal Type I error rate α. To check whether the Type I error values were accurate, 95% Agresti-Coull confidence intervals were derived (Agresti & Coull, 1998). These confidence intervals are [.038; .065] for $\alpha = .05$ and [.005; .019] for $\alpha = .01$. To judge whether the power is adequate, we used Cohen's (1988, p. 56) rule of thumb, considering a power value of .80 to be sufficiently high.

For some conditions, due to memory capacity problems, ML estimation was not possible and maximum empirical likelihood (MEL) estimation (Van der Ark et al., 2011) was used instead. MEL only uses the elements of $\mathbf{n}$ that are non-zero, and only the corresponding elements of $\mathbf{m}$ are estimated. The elements of $\mathbf{m}$ that correspond to a zero frequency are fixed to zero. Because the data are processed in a more efficient way, MEL estimation needs considerably less memory space. The major part of the study was programmed in R (R Core Team, 2014) using the R-package *cmm* (Bergsma & Van der Ark, 2013) for estimating marginal models. For the ADF method for testing H_{02} and H_{03}, following the procedure as described by Maydeu-Olivares et al. (2010), the simulations were done in Mplus Version 6.1 (Muthén & Muthén, 2010). However, for this method the small nominal Type I error ($\alpha = .01$) condition could not be tested due to limitations in the used software. Therefore, the R-package lavaan 0.5-9 (Rosseel, 2012) was used for testing this condition for H_{02} and H_{03}.

2.4.2 Results

Tables 2.1 and 2.2 show the Type I error rate and the power for testing hypotheses H_{01}, H_{02}, and H_{03} using the available statistical tests and the marginal modelling approach for the different conditions. For the Type I error rate, proportions outside the 95% confidence interval of the nominal α level are printed in bold. Results computed using MEL rather than ML

Table 2.1: *Type I Error Rate and Power for Testing H_{01} and H_{02} Using the Available Statistical Tests and the Marginal Modelling Approach.*

	H_{01}			H_{02}		
	Feldt	ADF	MM	Feldt	ADF	MM
Condition	1965			1969		
	Type I Error Rate					
Standard case	.058	**.068**	.041	.045	.040‡	.053
Low reliability ($\rho_{\alpha_1} = \rho_{\alpha_2} = .70$)	.049	**.070**	.045	.050	.053‡	.050
High reliability ($\rho_{\alpha_1} = \rho_{\alpha_2} = .90$)	.062	.062	**.066**	**.066**	.038‡	.054
Very high reliability ($\rho_{\alpha_1} = \rho_{\alpha_2} = .95$)	**.129**	.056	.053	**.123**	**.034**‡	.043
Polytomous items ($k = 4$)	.045	.054	*.050*	.044	.039	*.053*
More items ($J = 10$)	**.035**	.060	*.042*	.051	.040‡	*.058*
Small sample ($N = 100$)	.058	**.078**	.048	.041	.052‡	.054
Medium sample ($N = 500$)	.048	.054	.057	.038	.059‡	.053
Large sample ($N = 1000$)	.052	.060	.051	.039	.051‡	.052
Small nominal Type I error ($\alpha = .01$)	.007	**.030**	.005	.008	.007	.013
	Power					
Standard case	.984	.990	.982	.718	.691‡	.726
Small effect ($\rho_{\alpha_1} = .81$; $\rho_{\alpha_2} = .80$)	.112	.151	.124	.063	.056‡	.068
High reliability ($\rho_{\alpha_1} = .90$; $\rho_{\alpha_2} = .80$)	1.000	1.000	1.000	.992	.991‡	.991
Polytomous items ($k = 4$)	.976	.988	*.984*	.707	.753	*.755*
More items ($J = 10$)	.988	.987	*.991*	.777	.807‡	*.834*
Small sample ($N = 100$)	.845	.905	.842	.439	.407‡	.421
Medium sample ($N = 500$)	1.000	1.000	1.000	.989	.980‡	.987
Large sample ($N = 1000$)	1.000	1.000	1.000	1.000	1.000‡	1.000
Small nominal Type I error ($\alpha = .01$)	.937	.955	.923	.479	.474	.508

Note: ADF = asymptotic distribution free; MM = marginal modelling. The 95% CI for the Type I error rate equals [.038; .065] for $\alpha = .05$, and [.005; .019] for $\alpha = .01$. Values outside the 95% CI are printed in bold. Values computed using MEL are printed in italics. Values marked with a double dagger (‡) are based on less than 1,000 replications (convergence number between 885 and 996).

estimation are printed in italics. Values marked with a double dagger are based on fewer than 1,000 replications. For some replications (ranging from 4 to 115 per cell), the ADF method broke down.

For testing H_{01}, the marginal modelling approach yielded accurate Type I error rates, whereas Feldt's procedure was too liberal when reliability was

Table 2.2: *Type I Error Rate and Power for Testing H_{03} Using the Available Statistical Tests and the Marginal Modelling Approach.*

	H_{03}				
	Feldt-1980		AF	ADF	MM
Condition	Pitman	Δ			
	Type I Error Rate				
Standard case	**.070**	**.076**	.050	.042‡	.046
Low reliability ($\rho_{\alpha_1} = \rho_{\alpha_2} = .70$)	**.086**	**.086**	.054	.059‡	.039
High reliability ($\rho_{\alpha_1} = \rho_{\alpha_2} = .90$)	**.127**	**.124**	**.083**	.050‡	.048
Very high reliability ($\rho_{\alpha_1} = \rho_{\alpha_2} = .95$)	**.118**	**.148**	**.095**	.051‡	.064
Polytomous items ($k = 4$)	**.186**	**.158**	**.080**	.059	*.053*
More items ($J = 10$)	**.093**	**.095**	.059	.052‡	*.050*
Small sample ($N = 100$)	**.070**	**.066**	.054	.046‡	**.067**
Medium sample ($N = 500$)	**.083**	**.072**	.057	.052‡	.051
Large sample ($N = 1000$)	**.084**	**.071**	.046	.058‡	.051
Small nominal Type I error ($\alpha = .01$)	**.024**	.013	.010	.006	.005
	Power				
Standard case	.788	.796	.730	.715‡	.726
Small effect ($\rho_{\alpha_1} = .81$; $\rho_{\alpha_2} = .80$)	.091	.074	.055	.060‡	.063
High reliability ($\rho_{\alpha_1} = .90$; $\rho_{\alpha_2} = .80$)	.997	.998	.994	.994‡	.992
Polytomous items ($k = 4$)	.949	.939	.877	.871	*.906*
More items ($J = 10$)	.952	.952	.936	.928‡	*.924*
Small sample ($N = 100$)	.506	.537	.431	.419‡	.671
Medium sample ($N = 500$)	.988	.990	.981	.984‡	.999
Large sample ($N = 1000$)	1.000	1.000	1.000	1.000‡	1.000
Small nominal Type I error ($\alpha = .01$)	.611	.614	.499	.457	.483

Note: Δ = Delta method; AF = Alsawalmeh and Feldt (1994) procedure; ADF = asymptotic distribution free; MM = marginal modelling. The 95% CI for the Type I error rate equals [.038; .065] for $\alpha = .05$, and [.005; .019] for $\alpha = .01$. Values outside the 95% CI are printed in bold. Values computed using MEL are printed in italics. Values marked with a double dagger (‡) are based on less than 1,000 replications (convergence number between 885 and 996).

very high, and the ADF method was too liberal for a small sample size and a small nominal Type I error rate. Type I error rates just outside the 95% confidence interval were not interpreted. Except for the small effect condition, the three methods had similar adequate power.

For testing H_{02}, the marginal modelling approach and the ADF method yielded accurate Type I error rates, whereas Feldt's procedure was too liberal when reliability was very high. The three methods had similar power. The methods for testing H_{02} were less powerful than the methods for testing H_{01}. For conditions that are well known to reduce power (Cohen, 1988) — small sample, low nominal Type I error rate, and small effect — the power was especially low.

For testing H_{03}, the marginal modelling approach and the ADF method yielded accurate Type I error rates, whereas Feldt's procedures were generally too liberal (see Table 2.2). The method proposed by Alsawalmeh and Feldt (1994) was too liberal for polytomous items, high reliability, and very high reliability. With respect to power, the findings for hypothesis H_{03} were similar to the results found for H_{02} in most conditions. However, for the small sample condition the marginal modelling approach showed better results.

2.4.3 Discussion

The results of the simulation study showed that the marginal modelling approach generally resulted in accurate Type I error rates. The ADF method performed almost equally well but had poorer Type I error rates for small samples and small nominal Type I error rates for H_{01}. With respect to the small nominal Type I error rate, it seems that the tails of the distribution of r_α are not accurately estimated using the ADF procedure. With respect to small samples, Maydeu-Olivares et al. (2007) also found this result. An additional disadvantage is that for some data sets, the ADF method did not work. For testing alphas in dependent samples (H_{03}), Feldt's (1980) procedures had inaccurate Type I error rates in all conditions, suggesting that these tests are not robust against violations of the assumptions. We recommend not to use these tests in practical research.

The statistical tests for testing H_{01} had more power than those for H_{02} and H_{03}, which is due to H_{01} being a one-sided hypothesis and H_{02} and H_{03}

being two-sided hypotheses. Statistical tests for the same hypothesis had similar power. However, it may be noted that the power of a test can only be interpreted meaningfully if the Type I error is accurate; power and Type I error rate are usually a trade-off, and one can construct a very powerful test by always rejecting the null hypothesis. Hence, the power of tests having an inaccurate Type I error, such as the methods proposed by Feldt (1980), should be ignored.

2.5 General Discussion

This chapter features two innovations: the suggestion to use marginal models for testing hypotheses related to Cronbach's alpha; and the use of a data generation model for simulation studies that produces the desired population value of Cronbach's alpha and generates discrete data sets.

The marginal modelling approach was found to be more accurate than most of the available methods. It is very flexible because it is based on weak assumptions and can be generalized to more than two groups, to coefficients other than Cronbach's alpha, and to combinations of the hypotheses discussed in this chapter. These generalizations require adjusting the design matrices or constructing new design matrices. These generalizations are topics for future research. Outside the framework of marginal modelling such generalizations have been proposed. For instance, Hakstian and Whalen (1976), Bonett (2003), and Kim and Feldt (2008) generalized Feldt's (1969) method for testing H_{02} to more than two groups, and Woodruff and Feldt (1986) generalized Feldt's (1980) method for testing H_{03} to more than two groups. Kraemer (1981) extended H_{02} and H_{03} by proposing a test for the equality of two or more intraclass correlation coefficients.

The marginal modelling approach used to test the three hypotheses can also be used to construct a confidence interval for ρ_α, for $\rho_{\alpha_{g_1}} - \rho_{\alpha_{g_2}}$ from independent samples, and for $\rho_{\alpha_{t_1}} - \rho_{\alpha_{t_2}}$ from dependent samples. Wald confidence intervals for the three parameters can be obtained using the delta method. Let $g(\mathbf{n})$ equal a scalar sample statistic, for example r_α as expressed by the right-hand side of Equation 2.11, and let $\mathbf{G}(\mathbf{n})$ be the vector of first-order partial derivatives of $g(\mathbf{n})$ to $\mathbf{n}$. Under the assumption that $\mathbf{n}$ follows a multinomial distribution with covariance matrix $\mathbf{V}$, the asymptotic variance

of $g(\mathbf{n})$ equals $\mathbf{G}(\mathbf{n})\mathbf{V}\mathbf{G}(\mathbf{n})^T$. Its square root is the asymptotic standard error of r_α, from which the Wald confidence interval is constructed. For details on this method, we refer to Kuijpers et al. (2013b). Likelihood confidence intervals (for details, see Lang, 2008) for ρ_α can be constructed by testing the hypothesis H_{01} for a sequence of values of criterion c. The two values of c that result in p-values of .025 and .975, respectively, are the limits of the 95% likelihood confidence interval for ρ_α. The likelihood confidence interval is range preserving.

A limitation of the marginal modelling approach is that it requires much memory space, especially for a large number of dichotomously scored items, or for a medium-sized set of polytomously scored items. Due to this memory capacity problem, not all simulations could be done using ML estimation. Furthermore, marginal modelling needs much computation time for larger sets of items. To overcome these limitations, we recommend using the maximum empirical likelihood (MEL) method (Owen, 2001) that is implemented in a newer version of the *cmm* package. Initial simulation studies (Van der Ark et al., 2011) showed that ML and MEL produce similar results. Also in this study, there was no indication that the use of ML or MEL affected the results.

Our simulations showed that the ADF method (Maydeu-Olivares et al., 2007; Maydeu-Olivares et al., 2010) was accurate in most conditions; only for hypothesis H_{01} was the method too liberal, especially for a small sample size and a small nominal Type I error rate. However, the method has some practical limitations pertaining to the available software. First, if the data contain a dichotomous item with item mean equal to .50, then the MLM option (maximum likelihood with robust standard errors and a mean-adjusted chi-square test statistic) in Mplus (Muthén & Muthén, 2010) provides neither standard errors for any of the parameter estimates nor other fit indices. Consequently, the required standard errors of Cronbach's alpha could not be computed. If standard maximum likelihood estimation is used, this problem does not occur but then the estimated Type I error rates are poor because nonnormality is not taken into account. Second, for some samples (up to 12% in our replications) MLM estimation in Mplus breaks down, which might be due to the aforementioned problem. Third, for H_{02} and H_{03}, Mplus only

allows a nominal Type I error rate of .05. As a result, for our small nominal Type I error rate condition, we had to resort to the R-package *lavaan* 0.5-9 (Rosseel, 2012). The package *lavaan* reported NaNs (not a number) for standard errors of dichotomous items having a mean equal to .50, when using MLM estimation. However, *lavaan* produced a standard error for Cronbach's alpha and for the difference between the two alphas, but it is unclear whether or how the NaNs are taken into account. Fourth, the syntax of the ADF method in both Mplus and *lavaan* becomes large and laborious when the number of items exceeds ten. Because of these limitations, one has to be careful when using the ADF method, and further research is needed to solve these problems.

The other innovation was the data generation model in the simulation study. Because virtually all psychological tests produce discrete item scores, this way of simulating data is more realistic than in previous simulation studies, where continuous item scores were sampled from the moments of a continuous distribution. One may argue that the way we simulated data may favor the marginal modelling approach because the simulated data satisfy the assumptions of a marginal model but not the assumptions of the available statistical tests. However, the assumptions of a marginal model (multinomial distribution of the data) are so weak that almost any discrete data set satisfies the assumptions. So we consider this an advantage of the marginal modelling approach rather than a disadvantage of the simulation study.

Chapter 3

Standard Errors and Confidence Intervals for Scalability Coefficients in Mokken Scale Analysis Using Marginal Models

Abstract Mokken scale analysis is a popular method for scaling dichotomous and polytomous items. Whether or not items form a scale is determined by three types of scalability coefficients: (1) for pairs of items, (2) for items, and (3) for the entire scale. It has become standard practice to interpret the sample values of these scalability coefficients using Mokken's guidelines, which have been available since the 1970s. For valid assessment of the scalability coefficients, the standard errors of the scalability coefficients must be taken into account. So far, standard errors were not available for scales consisting of Likert items, the most popular item type in sociology, and standard errors could only be computed for dichotomous items if the number of items was small. This study solves these two problems. First, we derived standard errors for Mokken's scalability coefficients using a marginal modelling framework. These standard errors can be computed for all types of items used in

This chapter has been published as Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43,* 42-69.

Mokken scale analysis. Second, we proved that the method can be applied to scales consisting of large numbers of items. Third, we applied Mokken scale analysis to a set of polytomous items measuring tolerance. The analysis showed that ignoring standard errors of scalability coefficients might result in incorrect inferences.

3.1 Introduction

In the social sciences, researchers often use surveys or questionnaires for measuring the trait or attitude of interest, such as religiosity, tolerance or social capital. Typically, respondents react to a set of indicators of the trait. The indicators are generally referred to as items, and a set of items pertaining to the same trait is referred to as a scale. The respondents receive a score on each item. A summary of a respondent's item scores, most often the sum of the item scores, produces an estimate of his or her trait level. The sums of the item scores can only be used meaningfully as estimates of the respondents' trait levels if the scores on the items in the scale are unidimensional and have discrimination power to distinguish trait levels. Mokken scale analysis (Mokken, 1971; Sijtsma & Molenaar, 2002) is a popular method that can be used to partition a set of items into one or more unidimensional scales, possibly leaving some items unscalable. Some recent sociological studies that used Mokken scale analysis to construct scales investigated topics such as opinions on genetically modified foods (Loner, 2008), religious and spiritual beliefs (Gow, Watson, Whiteman, & Deary, 2011), political knowledge and media use (Hendriks Vettehen, Hagemann, & Van Snippenburg, 2004), social capital (Webber & Huxley, 2007), and attitudes toward illegal immigration (Ommundsen, Mörch, Hak, Larsen, & Van der Veer, 2002).

In Mokken scale analysis, three types of scalability coefficients are used both as criteria for the item partitioning and as diagnostics for the strength of the scales: (1) H_{ij}, a coefficient for the scalability of item pair (i, j); (2) H_j, a coefficient for the scalability of item j; and (3) H, a coefficient for the scalability of the entire scale. Details of the scalability coefficients are discussed in Section 3.2 of this chapter. Mokken (1971, p. 184) advocated

that items form a scale if, and only if,

$$\rho_{ij} > 0 \text{ (which is equivalent to } H_{ij} \geq 0) \text{ for all } i < j, \text{ and} \tag{3.1}$$

$$H_j \geq c \text{ for all } j, \tag{3.2}$$

where ρ is the product-moment correlation and c some positive lower bound specified by the researcher. He proposed to choose the lower bound c to be at least equal to .3, in order to keep nondiscriminating items and weakly discriminating items out of the scale (Sijtsma & Molenaar, 2002). He also advocated that H should be at least .3 and he considered a scale to be weakly scalable if $.3 \leq H < .4$, moderately scalable if $.4 \leq H < .5$, and strongly scalable if $H \geq .5$ (Mokken, 1971, p. 185), whereas $H < .3$ meant that the items are unscalable. For example, for the 6-item scale *Personal Skills* ($N = 279$), Webber and Huxley (2007) found that all H_{ij}s were positive, the values of H_j ranged between .32 and .45, and $H = .37$. They concluded that *Personal Skills* had "sufficient scale H values to be useful". We argue that researchers should take into account the uncertainty of the estimated scalability coefficients when applying Mokken's heuristic guidelines. The uncertainty is quantified by the standard errors of the estimated values. If the standard error of H is small, then Webber and Huxley's conclusion is justified, but if the standard error is large (for example, .08) then there is a reasonable chance that the population value of H is less than .3, and that the set of items that constitute *Personal Skills* is in fact unscalable following Mokken's guidelines. A similar line of reasoning applies when H_{ij} and H_j are evaluated.

Although some studies derived standard errors for scalability coefficients, none yielded standard errors for all scalability coefficients that could also be applied to reasonable or large numbers of items. Mokken (1971, pp. 164-169) derived asymptotic standard errors of H in the case of dichotomous items. Van Onna (2004) used several computer-intensive methods to compute confidence intervals for coefficient H, both for dichotomous and polytomous items, and advocated using the nonparametric bootstrap for computing a range-preserving confidence interval for H. Van der Ark, Croon, and Sijtsma (2008a) used marginal modelling as a framework for testing specific hypotheses about scalability coefficients H_{ij}, H_j, and H. Within this framework they

also derived standard errors for H_{ij}, H_j, and H. However, their approach could be applied only to small sets of dichotomous items. A practical problem is that none of the methods have been implemented in software, which makes the methods unavailable for applied researchers. As a result, standard errors of scalability coefficients are never reported in applications of Mokken scale analysis.

In this chapter, we solve all limitations mentioned. We generalize the marginal modelling approach for computing standard errors of scalability coefficients to polytomous items and to large numbers of items. Furthermore, the approach is made available in the software package *mokken* (Van der Ark, 2007, 2012). The remainder of this chapter is organized as follows. First, we discuss Mokken scale analysis. Second, we discuss the general principle of obtaining standard errors of sample statistics using the marginal modelling approach, we give detailed results for the derivation of standard errors of scalability coefficients for polytomous items, and we discuss how the method can be applied to large numbers of items. Third, we estimate the scalability coefficients and their standard errors for two real-data examples. The examples demonstrate that ignoring the uncertainty of the estimated scalability coefficients may lead to incorrect inferences. Finally, we discuss the strengths and weaknesses of the approach.

3.2 Mokken Scale Analysis

3.2.1 The Monotone Homogeneity Model

Mokken scale analysis is based on the monotone homogeneity model (Mokken, 1971, Chapter 4; Sijtsma & Molenaar, 2002, pp. 22-23), which is a nonparametric item response theory (IRT) model for measuring respondents on an ordinal scale. We consider a set of J items numbered $1, 2, \ldots, J$, each having $z + 1$ ordered answer categories $x = 0, 1, \ldots, z$. Let X_j denote the score on item j and let $X_+ = \sum_j X_j$ denote the sum of the J item scores. Let θ denote a possibly multidimensional latent variable (usually referred to as *latent trait*); often θ values are interpreted in terms of the construct that the items measure in common. IRT models describe the relation between latent trait θ and the probabilities of item scores x, $P(X_j = x|\theta)$. The monotone

homogeneity model consists of three assumptions:

Unidimensionality: The latent variable θ is unidimensional;

Local independence: The item scores are independent given θ; that is,

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_J = x_J|\theta) = \prod_{j=1}^{J} P(X_j = x_j|\theta).$$

Monotonicity: The probability of having a score of at least x on item j, $P(X_j \geq x|\theta)$, is a nondecreasing function of θ.

The monotone homogeneity model is a general model in the sense that all other popular unidimensional IRT models are a special case of the monotone homogeneity model (Van der Ark, 2001). For practical purposes, the model allows the stochastic ordering of θ by means of X_+ (for details, see Van der Ark & Bergsma, 2010, and references therein). Hence, only if the monotone homogeneity model fits the data well, the total scale score can be used meaningfully to order respondents.

Mokken scale analysis can be regarded as a set of methods to construct scales for which the monotone homogeneity model and other nonparametric IRT models fit well. The general idea is that one investigates observable properties implied by the model. For example, under the monotone homogeneity model all scalability coefficients H_{ij} must be nonnegative. Hence, if a researcher finds that for a particular scale the sample values of H_{ij} are all nonnegative, then this result supports the possibility that the monotone homogeneity model is true, whereas negative H_{ij} values mean that the model must be rejected.

3.2.2 Scalability Coefficients

Item Steps and Weighted Guttman Errors

Scalability coefficients H_{ij}, H_j, and H are based on item steps and Guttman errors (Molenaar, 1991), which are best explained by means of an example. Table 3.1 (see Weijmar Schultz & Van der Wiel, 1991) shows a cross-classification of the scores of $N = 178$ respondents on $J = 2$ items (Item a and Item b), each having $z + 1 = 4$ ordered answer categories. The frequencies are denoted n_{ab}^{xy} $x, y = 0, \ldots, 3$, and the marginal frequencies are denoted n_{ab}^{x+} and n_{ab}^{+y}, where the "+" indicates the sum over all categories.

Table 3.1: *Cross-Tabulation of Scores on Item a and Item b for N=178 Respondents; Guttman Weights Are Between Parentheses.*

	X_b					
X_a	0	1	2	3	n_{ab}^{x+}	$P(X_a \geq x)$
0	**3** (0)	0 (2)	0 (4)	0 (7)	3	1.000
1	**4** (0)	7 (1)	3 (2)	0 (4)	14	.983
2	**10** (0)	**22** (0)	**34** (0)	3 (1)	69	.904
3	9 (2)	17 (1)	**40** (0)	**26** (0)	92	.517
n_{ab}^{+y}	26	46	77	29	178	
$P(X_b \geq y)$	1.000	.854	.596	.163		

Note: Frequencies of response patterns that are not Guttman errors are printed bold.

Item steps are boolean statements $X_j \geq x$ $(j = 1, \ldots, J; x = 0, \ldots, z)$, indicating whether a respondent has passed the item step $(X_j \geq x)$ or not $(X_j < x)$. The popularity of an item step is determined by means of the proportion of respondents that has passed the item step, $P(X_j \geq x)$. It may be noted that $P(X_j \geq 0) = 1$ by definition, and this probability thus is not informative. The ordering of the $2z$ item steps in Table 3.1 by descending popularity equals

$$X_a \geq 1, X_a \geq 2, X_b \geq 1, X_b \geq 2, X_a \geq 3, X_b \geq 3. \tag{3.3}$$

Respondents who did not pass any item step have item-score pattern $(0, 0)$; respondents who have passed one item step, most likely have passed the most popular item step $X_a \geq 1$, producing item-score pattern $(1, 0)$; respondents who have passed two item steps, most likely have passed $X_a \geq 1$ and $X_a \geq 2$, producing item-score pattern $(2, 0)$, and so on. The admissable item-score patterns are (0,0), (1,0), (2,0), (2,1), (2,2), (3,2), and (3,3) (frequencies printed in bold in Table 3.1) that are consistent with the order of the item steps. Each respondent that passes the h most popular item steps and does not take the remaining $2z - h$ less popular item steps has an item-score pattern that is in agreement with the Guttman (1950) model (Molenaar, 1991). Such admissable patterns are called *conformal patterns*. Respondents having item-score pattern (0,3) passed the least popular item step $X_b \geq 3$

but did not pass the more popular item steps $X_a \geq 1$, $X_a \geq 2$, and $X_a \geq 3$. Patterns for which at least one less popular item step has been passed and one more popular has not been passed are called *Guttman errors* (Molenaar, 1991). A set of items is perfectly scalable if there are no Guttman errors, and is less scalable as the number of Guttman errors increases.

Molenaar (1991) suggested weighting the frequencies of the Guttman errors depending on the degree of deviation from item-score patterns yielding a perfect scale. The weight for the frequency of a particular item-score pattern is computed as follows. We consider all pairs of item steps and we compute the weight equal to the number of pairs of item steps for which the less popular item step was passed and the more popular step was failed. For example, for item-score pattern (0,2) in Table 3.1, the Guttman weight equals $w_{ab}^{02} = 4$ because for four pairs of item steps $(X_a \geq 1, X_b \geq 1)$, $(X_a \geq 1, X_b \geq 2)$, $(X_a \geq 2, X_b \geq 1)$, and $(X_a \geq 2, X_b \geq 2)$ the less popular item step was passed and the more popular step was failed — for example, for pair $(X_a \geq 1, X_b \geq 1)$, the less popular item step $X_b \geq 1$ was passed but the more popular item step $X_a \geq 1$ was failed. The weights are shown between parentheses in each cell of Table 3.1. It may be noted that the boldfaced conformal item-score patterns have a weight equal to zero.

For computational purposes, we give a formula for computing the weights (see also Ligtvoet, Van der Ark, Te Marvelde, & Sijtsma, 2010). Let the $2z$ item steps be ordered by descending popularity (cf. Equation 3.3), and let $\mathbf{q}_{ij}^{xy} = (q_{ij(1)}^{xy}, q_{ij(2)}^{xy}, \ldots, q_{ij(2z)}^{xy})$ be a vector consisting of zeroes and ones indicating for item-score pattern $(X_i = x, X_j = y)$ whether an item step has been passed (1) or not (0). Then weight w_{ij}^{xy} equals

$$w_{ij}^{xy} = \sum_{u=2}^{2z} q_{ij(u)}^{xy} \left(\sum_{v=1}^{u-1} |1 - q_{ij(v)}^{xy}| \right). \tag{3.4}$$

Equation 3.4 counts how often a score 0 precedes a score 1 in vector $\mathbf{q}_{ij}^{xy}$. For example, it may be noted that for response pattern (0,2) in Table 3.1, the third and fourth item steps in Equation 3.3 are passed, and so $\mathbf{q}_{ab}^{02} = (0, 0, 1, 1, 0, 0)$. In $\mathbf{q}_{ab}^{02}$, the score 0 precedes the score 1 four times, and so the weight w_{ab}^{02} equals 4. As a second example, consider the item-score pattern (2,1). Here, the first, second, and third item steps are passed, and thus

$\mathbf{q}_{ab}^{21} = (1, 1, 1, 0, 0, 0)$. Here, there are no occasions on which a score 0 precedes a score 1, and thus the weight w_{ab}^{21} is equal to 0.

Item Pair Scalability Coefficients

Item pair scalability coefficient H_{ij} compares the sum of weighted observed frequencies of Guttman errors to the sum of weighted frequencies of Guttman errors that is expected under marginal independence of the item scores. Let

$$e_{ij}^{xy} = \frac{n_{ij}^{x+} \times n_{ij}^{+y}}{N} \tag{3.5}$$

be the expected bivariate frequency under marginal independence; let F_{ij} and E_{ij} be the sum of weighted observed and expected frequencies of Guttman errors, respectively, for item pair (i, j). Then

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}} = 1 - \frac{\sum_x \sum_y w_{ij}^{xy} n_{ij}^{xy}}{\sum_x \sum_y w_{ij}^{xy} e_{ij}^{xy}}. \tag{3.6}$$

If there are no Guttman errors, then $H_{ij} = 1$; if there are as many Guttman errors as there are under marginal independence, then $H_{ij} = 0$. Under the monotone homogeneity model, $H_{ij} \geq 0$. Molenaar (1991) showed that H_{ij} can be written as a normed covariance. Let σ_{ij} be the covariance between item i and item j and let $\sigma_{ij}^{\max}$ be the maximum covariance between item i and item j, given the marginal distributions of both items. Given that the items both have a positive variance, $H_{ij} = \sigma_{ij}/\sigma_{ij}^{\max}$. For a set of J items, let $K = \frac{1}{2}J(J-1)$ denote the number of item pairs; hence, we have K different coefficients H_{ij}.

The Item Scalability Coefficient

Item scalability coefficient H_j is a generalization of H_{ij}; it compares the sum of weighted observed and weighted expected frequencies of Guttman errors for an individual item:

$$H_j = 1 - \frac{\sum_{i \neq j} F_{ij}}{\sum_{i \neq j} E_{ij}} = 1 - \frac{\sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} n_{ij}^{xy}}{\sum_{i \neq j} \sum_x \sum_y w_{ij}^{xy} e_{ij}^{xy}}. \tag{3.7}$$

Under the monotone homogeneity model, $0 \leq H_j \leq 1$. Let $R_{(j)} = X_+ - X_j$ denote the *rest score*. Sijtsma and Molenaar (2002, p. 57) showed that

H_j is equal to the normed covariance between X_j and $R_{(j)}$; that is, $H_j = \sigma_{jR_{(j)}}/\sigma^{\max}_{jR_{(j)}}$. Hence, H_j expresses the strength of the association between item j and the other items in the scale, and it can be viewed as the nonparametric analogue of the discrimination parameter in parametric IRT (e.g., Van Abswoude, Van der Ark, & Sijtsma, 2004). To keep nondiscriminating items and weakly discriminating items out of the scale, Mokken (1971, p. 184) proposed that all H_js should be greater than some lower bound $c > 0$. It may be noted that $c > 0$ is not an observable property of the monotone homogeneity model.

The Total-Scale Scalability Coefficient

Coefficient H is a generalization of H_{ij} and H_j; it compares the sum of weighted observed and weighted expected frequencies of Guttman errors for all J items in the entire scale:

$$H = 1 - \frac{\sum\sum_{i\neq j} F_{ij}}{\sum\sum_{i\neq j} E_{ij}} = 1 - \frac{\sum\sum_{i\neq j}\sum_x\sum_y w_{ij}^{xy} n_{ij}^{xy}}{\sum\sum_{i\neq j}\sum_x\sum_y w_{ij}^{xy} e_{ij}^{xy}}. \tag{3.8}$$

H expresses the scalability of all items in the scale. Under the monotone homogeneity model, $0 \leq H \leq 1$. Moreover, Mokken (1971, pp. 148-153; see also Sijtsma & Molenaar, 2002, Theorem 4.2) showed that under the monotone homogeneity model, the scalability coefficients are related in such a way that

$$\min_{i,j}(H_{ij}) \leq \min_j(H_j) \leq H \leq \max_j(H_j) \leq \max_{i,j}(H_{ij}).$$

3.2.3 Methods in Mokken Scale Analysis

Mokken scale analysis contains an automated item selection procedure that partitions the set of items into one or more unidimensional scales. A scale is considered a Mokken scale if it satisfies the two criteria as stated in Equations 3.1 and 3.2. Moreover, Mokken scale analysis provides several methods for the additional investigation of the assumptions of the monotone homogeneity model and other nonparametric IRT models. A description of these methods is beyond the scope of this chapter, and we refer the interested reader to, for example, Mokken (1971) and Sijtsma and Molenaar (2002).

3.3 Standard Errors of Scalability Coefficients

In marginal modelling of categorical data (e.g., see Bergsma et al., 2009, and references therein), a two-step method is used to compute standard errors of sample statistics. We describe this method for the scalability coefficients. The first step is to write the scalability coefficients as a function of the frequencies of the observed item-score patterns in the data. A set of J items, each with $z+1$ ordered answer categories $(0, 1, \ldots, z)$ produces $L = (z+1)^J$ possible item-score patterns. Without loss of generality, we assume that item-score patterns are in lexicographic order: going from $00\ldots0$ to $zz\ldots z$ with the last digit changing fastest, and the digit in the first column changing slowest. The observed frequencies of the L possible item-score patterns can be collected in a vector $\mathbf{n}$. For example, a set of $J = 3$ items (denoted by a, b, and c) each with $(z+1) = 3$ answer categories has $L = 3^3 = 27$ possible item-score patterns; hence vector $\mathbf{n}$ equals

$$\mathbf{n} = \begin{pmatrix} n_{abc}^{000} \\ n_{abc}^{001} \\ n_{abc}^{002} \\ n_{abc}^{010} \\ n_{abc}^{011} \\ \vdots \\ n_{abc}^{220} \\ n_{abc}^{221} \\ n_{abc}^{222} \end{pmatrix}. \tag{3.9}$$

Vector $\mathbf{n}$ in Equation 3.9 is used throughout to illustrate the approach. Let vector $\mathbf{H}_{ij} = (H_{12}, H_{13}, \ldots, H_{J-1,J})^T$ (the superscript T denotes the transpose) contain all K scalability coefficients H_{ij}, and let vector $\mathbf{H}_j = (H_1, H_2, \ldots, H_J)^T$ contain all J scalability coefficients H_j. Also, let $\mathbf{g}$ and $\mathbf{g}^\dagger$ be vector-valued functions, and let $g^\ddagger$ be a scalar function. We show that the scalability coefficients can be written as a function of $\mathbf{n}$; that is

$$\mathbf{H}_{ij} = \mathbf{g}(\mathbf{n}) \tag{3.10}$$

$$\mathbf{H}_j = \mathbf{g}^\dagger(\mathbf{n}) \tag{3.11}$$

$$H = g^\ddagger(\mathbf{n}) \tag{3.12}$$

The second step is to use the delta method to obtain the asymptotic standard errors for the scalability coefficients. Let $\mathbf{V_n}$ and $\mathbf{V_{g(n)}}$ be the asymptotic variance-covariance matrix of $\mathbf{n}$ and $\mathbf{g}(\mathbf{n})$, respectively; let N be the total sample size; and let $\mathbf{D}(\mathbf{x})$ be a diagonal matrix with the elements of vector $\mathbf{x}$ on the diagonal.

If $\mathbf{n}$ is sampled from a multinomial distribution, then

$$\mathbf{V_n} = \mathbf{D}(\mathbf{n}) - \mathbf{n}N^{-1}\mathbf{n}^T$$

(e.g., Agresti, 2007, p. 6). Now if $\mathbf{G} = \mathbf{G}(\mathbf{n})$ is the Jacobian, which is the matrix of first partial derivatives of $\mathbf{g}(\mathbf{n})$ to $\mathbf{n}$, then according to the delta method

$$\begin{aligned}\mathbf{V_{g(n)}} &= \mathbf{G}\mathbf{V_n}\mathbf{G}^T \\ &= \mathbf{G}\left[\mathbf{D}(\mathbf{n}) - \mathbf{n}N^{-1}\mathbf{n}^T\right]\mathbf{G}^T \\ &= \mathbf{G}\mathbf{D}(\mathbf{n})\mathbf{G}^T - \mathbf{G}\mathbf{n}N^{-1}\mathbf{n}^T\mathbf{G}^T. \end{aligned} \tag{3.13}$$

In most applications of marginal models, the functions $\mathbf{g}(.)$ are homogeneous of order 0; that is, the value of $\mathbf{g}(.)$ does not change when the values of its arguments are all multiplied by the same constant t:

$$\mathbf{g}(t\mathbf{n}) = \mathbf{g}(\mathbf{n}).$$

For such functions it does not matter whether $\mathbf{n}$ represents the observed frequencies or the observed probabilities. Functions $\mathbf{g}(\mathbf{n})$ (Equation 3.10), $\mathbf{g}^\dagger(\mathbf{n})$ (Equation 3.11), and $g^\ddagger(\mathbf{n})$ (Equation 3.12) are also homogeneous functions. Euler's homogeneous function theorem (e.g., Weisstein, 2011) now implies that $\mathbf{G}\mathbf{n} = \mathbf{0}$. As a result, Equation 3.13 reduces to

$$\mathbf{V_{g(n)}} = \mathbf{G}\mathbf{D}(\mathbf{n})\mathbf{G}^T. \tag{3.14}$$

Taking the square root of the diagonal of $\mathbf{V_{g(n)}}$ produces the required standard errors.

We demonstrate how to obtain $\mathbf{g}(\cdot)$ (Equation 3.10), $\mathbf{g}^\dagger(\cdot)$ (Equation 3.11), and $g^\ddagger(\cdot)$ (Equation 3.12). The notation used in these derivations is called the *generalized exp-log notation* (Bergsma, 1997; Kritzer, 1977). Moreover, we also show how to obtain the matrix of first partial derivatives for these functions.

3.3.1 Generalized Exp-Log Notations for the Three Scalability Coefficients

Let $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$, and $\mathbf{A}_5$, be design matrices to be explained below. Matrix $\mathbf{A}_1$ is explained in detail to give the reader more insight into the generalized exp-log notation. The construction of the other design matrices is relegated to Appendix 3.A. The generalized exp-log notation for $\mathbf{H}_{ij}$ (Equation 3.10) is

$$\mathbf{H}_{ij} = \mathbf{g}(\mathbf{n}) = \mathbf{A}_5 \exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n})))). \tag{3.15}$$

The notation $\exp(\mathbf{X})$ and $\log(\mathbf{X})$ denote the exponential and logarithmic functions, evaluated element-wise to the elements of $\mathbf{X}$.

Let $\mathbf{n}_{ij}$ be the vector containing the $(z+1)^2$ bivariate frequencies of item pair (i, j). For K item pairs, the total number of bivariate frequencies equals $B = K(z+1)^2$. Let $\mathbf{n}_j$ be the vector containing the $(z+1)$ univariate frequencies of item (j). For J items the total number of univariate frequencies equals $U = J(z+1)$. For example, for Equation 3.9

$$\mathbf{n}_{ab} = \begin{pmatrix} n_{abc}^{00+} \\ n_{abc}^{01+} \\ n_{abc}^{02+} \\ n_{abc}^{10+} \\ n_{abc}^{11+} \\ n_{abc}^{12+} \\ n_{abc}^{20+} \\ n_{abc}^{21+} \\ n_{abc}^{22+} \end{pmatrix} \text{ and } \mathbf{n}_a = \begin{pmatrix} n_{abc}^{0++} \\ n_{abc}^{1++} \\ n_{abc}^{2++} \end{pmatrix}.$$

The $(B + U + 1) \times L$ design matrix $\mathbf{A}_1$ consists of three submatrices:

$$\mathbf{A}_1 = \begin{pmatrix} \mathbf{B} \\ \mathbf{U} \\ \mathbf{1}_L^T \end{pmatrix}. \tag{3.16}$$

The $B \times L$ submatrix $\mathbf{B}$ is necessary to obtain the B observed bivariate frequencies. The first $(z+1)^2$ rows correspond to the first item pair (item 1, item 2); the next $(z+1)^2$ rows correspond to the second item pair (item 1, item 3), and so on; the L columns correspond to the L item-score patterns. Element (b, l) equals 1 if the l-th item-score pattern contributes to the b-th

bivariate frequency, and element (b, l) equals 0 otherwise. For example, for the vector of observed frequencies in Equation 3.9, the first row of $\mathbf{B}$, which pertains to observed bivariate frequency $n_{abc}^{00+} = n_{abc}^{000} + n_{abc}^{001} + n_{abc}^{002}$, equals

$$(1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0).$$

The $U \times L$ submatrix $\mathbf{U}$ is necessary to obtain the U observed univariate frequencies. The first $(z+1)$ rows correspond to item 1; the next $(z+1)$ rows correspond to item 2, and so on. Element (u, l) equals 1 if the l-th item-score pattern contributes to the u-th observed univariate frequency, and element (u, l) equals 0 otherwise. For example, for the vector of observed frequencies in Equation 3.9, the first row of $\mathbf{U}$, which pertains to observed univariate frequency n_{abc}^{0++}, equals

$$(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0).$$

Vector $\mathbf{1}_L^T$ is the $1 \times L$ unit vector. For the vector of observed frequencies in Equation 3.9

$$\mathbf{A}_1 \cdot \mathbf{n} = \begin{pmatrix} \mathbf{B} \\ \mathbf{U} \\ \mathbf{1}_L^T \end{pmatrix} \cdot \mathbf{n} = \begin{pmatrix} \mathbf{n}_{ab} \\ \mathbf{n}_{ac} \\ \mathbf{n}_{bc} \\ \mathbf{n}_a \\ \mathbf{n}_b \\ \mathbf{n}_c \\ N \end{pmatrix}. \tag{3.17}$$

Design matrices $\mathbf{A}_2$, $\mathbf{A}_3$, $\mathbf{A}_4$, and $\mathbf{A}_5$ are constructed in a similar way (see Appendix 3.A).

The generalized exp-log notation for $\mathbf{H}_j$ (Equation 3.11) is

$$\mathbf{H}_j = \mathbf{g}^{\dagger}(\mathbf{n}) = \mathbf{A}_5^{\dagger} \exp(\mathbf{A}_4^{\dagger} \log(\mathbf{A}_3^{\dagger} \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n})))). \tag{3.18}$$

Note that $\mathbf{A}_1$ and $\mathbf{A}_2$ in Equation 3.18 are equal to those in Equation 3.15. Design matrices $\mathbf{A}_3^{\dagger}$, $\mathbf{A}_4^{\dagger}$, and $\mathbf{A}_5^{\dagger}$ are derived in Appendix 3.B.

The generalized exp-log notation for H (Equation 3.12) is

$$H = g^{\ddagger}(\mathbf{n}) = \mathbf{A}_5^{\ddagger} \exp(\mathbf{A}_4^{\ddagger} \log(\mathbf{A}_3^{\ddagger} \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n})))). \tag{3.19}$$

Note that $\mathbf{A}_1$ and $\mathbf{A}_2$ in Equation 3.19 are equal to those in Equation 3.15. Design matrices $\mathbf{A}_3^{\ddagger}$, $\mathbf{A}_4^{\ddagger}$, and $\mathbf{A}_5^{\ddagger}$ are derived in Appendix 3.C. Once the

design matrices have been constructed, the matrix of partial derivatives $\mathbf{G}$ can be derived (Appendix 3.D). Implementing $\mathbf{G}$ into Equation 3.14 produces the required standard errors.

3.3.2 Standard Errors for Scales Consisting of Large Numbers of Items

A practical problem that occurs is that the proposed method for deriving standard errors for scalability coefficients cannot be applied to large numbers of items (cf. Van der Ark et al., 2008a). Even for relatively small scales, L can be so large that vector $\mathbf{n}$ and the $(B+U+1) \times L$ matrix $\mathbf{A}_1$ (Equation 3.16) cannot be stored in computer memory. For large numbers of items, B may also be too large to store $\mathbf{A}_2$ and $\mathbf{A}_3$. For example, for $J = 10$ Likert items with $z + 1 = 5$ ordered answer categories, $L = 5^{10} = 9,765,625$ and $B = \binom{10}{2}5^2 = 1125$. Two modifications in the generalized exp-log notation reduce the computational burden considerably, so that standard errors of scalability coefficients can be computed for up to approximately 100 items and up to approximately 100,000 respondents. However, for larger data sets, computation may be slow.

The largest contribution to reducing the computational burden is to use only the non-zero frequencies in $\mathbf{n}$, which pertain to item-score patterns that are observed in the data, and collect them in vector $\mathbf{n}^*$. So, all elements of $\mathbf{n}^*$ are positive and the size of $\mathbf{n}^*$, denoted L^*, cannot exceed the sample size N. Let a matrix superscripted with an asterisk indicate a *reduced matrix*, which means that the rows and/or columns pertaining to zero-frequencies have been deleted. Thus, when only the non-zero observed frequencies are used, expression $\mathbf{A}_1.\mathbf{n}$ in Equations 3.15, 3.18, and 3.19 is replaced by $\mathbf{A}_1^*.\mathbf{n}^*$, and expression $\mathbf{GD}(\mathbf{n})\mathbf{G}^T$ is replaced by $\mathbf{G}^*\mathbf{D}(\mathbf{n}^*)\mathbf{G}^{*T}$. Other matrices used in this chapter remain unchanged. Typically, because L^* is much smaller than L, the reduced vectors and matrices are small enough to be stored in computer memory. We show that using reduced matrices does not affect the computation of the scalability coefficients and their standard errors.

First, we show that $\mathbf{A}_1.\mathbf{n} = \mathbf{A}_1^*.\mathbf{n}^*$, which means that Equations 3.15, 3.18, and 3.19 are unaffected by using reduced matrices.

Proof. Let $\sum_{l=1}^{L} A_{i,l} n_l$ be the i-th element in vector $\mathbf{A}_1.\mathbf{n}$. If $n_l = 0$ then

$A_{i,l}n_l$ has no contribution to the i-th element in $\mathbf{A}_1.\mathbf{n}$, and the l-th column of $\mathbf{A}_1$ and the l-th element of $\mathbf{n}$ can be removed without consequences. □

Second, we show that $\mathbf{GD}(\mathbf{n})\mathbf{G}^T = \mathbf{G}^*\mathbf{D}(\mathbf{n}^*)\mathbf{G}^{*T}$, which means that the computation of the standard errors in Equation 3.14 is unaffected by using reduced matrices.

Proof. Let $\mathbf{G}_l$ denote the l-th column of matrix $\mathbf{G}$. Hence, $\mathbf{GD}(\mathbf{n})\mathbf{G}^T = \sum_{l=1}^{L} \mathbf{G}_l\mathbf{G}_l^T n_l$. If $n_l = 0$ then $\mathbf{G}_l\mathbf{G}_l^T n_l = 0$; and neither the l-th column of $\mathbf{G}$ nor the l-th element of $\mathbf{n}$ have any contribution to $\mathbf{GD}(\mathbf{n})\mathbf{G}^T$ and can be removed without consequences. □

In general, direct computation of the design matrices $\mathbf{A}_1^*$, $\mathbf{A}_2$, and $\mathbf{A}_3$ is unnecessary and can be avoided, which is convenient when the number of observed bivariate frequencies B is large. The procedure is described in Appendix 3.D.

3.4 Mokken Scale Analysis of Data Measuring Tolerance

The use of marginal modelling for the derivation of standard errors and the accompanying confidence intervals is illustrated by means of data from the 2008 European Values Study (EVS, 2011). This large-scale cross-national survey provides insight into the basic values, preferences, attitudes and opinions that people all over Europe have about, for instance life, work, family, sexual behavior, gender roles, politics, religion, well-being, and tolerance. We analyze data pertaining to the tolerance scale. The tolerance scale consists of 20 items, where one part of the items measures tolerance with respect to material issues, and the other part measures tolerance with respect to interpersonal issues. Each item pertains to a particular controversial behavior, and the respondents had to indicate the degree to which they consider the behavior to be justified. Some examples are "Do you justify adultery?", "Do you justify euthanasia?", and "Do you justify prostitution?". In the original data set, the answer categories ranged from 1 (*never*) to 10 (*always*). The more extreme response categories were almost never chosen by respondents, and so the corresponding cell frequencies were close to or equal to zero. For this study, the answer categories were recoded into three categories, with the

scores 1 to 3 being recoded into 1, the scores 4 to 7 into 2, and scores 8 to 10 into 3.

Mokken scale analyses were performed on the data obtained in the Netherlands ($N = 1,554$), presumably a rather liberal country with respect to tolerance, and the former Soviet republic of Georgia ($N = 1,500$), presumably a rather conservative country (for the computer syntax, see Appendix 3.E). These two countries were chosen to show that in some cases standard errors do affect the conclusions, and in other cases they do not. Since no or almost no cases were in the third category, for the Georgian sample, two items (i.e, items 3 and 4) were deleted from the tolerance scale. Note that for the analyses we used the same items for both samples. However, the scales discussed hereunder are not identical.

For the Dutch sample, the automated item selection procedure (see Section 3.2.3) produced three scales, but only the first scale will be considered here. The first scale consisted of 12 items, and measured tolerance with respect to interpersonal issues. The following items were included in the scale: Do you justify ... taking soft drugs (item 4); adultery (item 6); homosexuality (item 8); abortion (item 9); divorce (item 10); euthanasia (item 11); suicide (item 12); having casual sex (item 14); avoiding a fare on public transport (item 15); prostitution (item 16); experiments on human embryos (item 17); and invitro fertilization (item 19).

Table 3.2 shows the sample values of H_{ij} and H_j plus their asymptotic standard errors for the first scale of the Dutch sample. To assess whether the item pair scalability coefficients were significantly greater than zero, 95% confidence intervals were obtained using $\widehat{H}_{ij} \pm 1,96 * se(\widehat{H}_{ij})$. Because the value zero was not included in the confidence interval for any of the 66 sample H_{ij}s, all $\widehat{H}_{ij}$s were significantly greater than zero. Similarly, 95% confidence intervals were created for H_j. Because the confidence interval included the criterion value $c = .3$ only for item 15 ($\widehat{H}_{15} = .303$; s.e. $= .024$), we do not have sufficient evidence that item 15 satisfies the second property of a Mokken scale (i.e., $H_j \geq c$ for all j) and thus it may be considered for removal from the scale. Following Mokken's guidelines, the items form a scale of moderate strength ($\widehat{H} = .479$; s.e. $= .012$).

For the Georgian sample, the automated item selection procedure pro-

Table 3.2: *Scalability Coefficients H_{ij} and H_j and Their Standard Errors (Between Brackets) for 12 Items Measuring Tolerance with Respect to Interpersonal Issues for the Dutch Sample.*

	H_{ij}											H_j
	4	6	8	9	10	11	12	14	15	16	17	
4: Soft Drugs												.436
												(.019)
6: Adultery	.325											.412
	(.036)											(.022)
8: Homosexuality	.652	.539										.584
	(.044)	(.060)										(.018)
9: Abortion	.432	.460	.662									.554
	(.033)	(.035)	(.024)									(.015)
10: Divorce	.518	.475	.733	.750								.573
	(.037)	(.040)	(.023)	(.021)								(.015)
11: Euthanasia	.502	.490	.588	.715	.635							.544
	(.040)	(.047)	(.027)	(.022)	(.024)							(.016)
12: Suicide	.437	.426	.596	.533	.524	.531						.448
	(.034)	(.037)	(.032)	(.028)	(.029)	(.025)						(.016)
14: Casual Sex	.585	.526	.643	.557	.544	.530	.471					.510
	(.029)	(.035)	(.037)	(.029)	(.027)	(.031)	(.029)					(.016)
15: Fare Public Transp.	.282	.234	.457	.313	.345	.367	.273	.398				.303
	(.036)	(.036)	(.062)	(.036)	(.040)	(.046)	(.036)	(.035)				(.024)
16: Prostitution	.458	.428	.522	.520	.639	.587	.455	.609	.336			.492
	(.032)	(.036)	(.029)	(.027)	(.026)	(.027)	(.028)	(.027)	(.036)			(.016)
17: Human Embryos	.297	.338	.414	.514	.429	.436	.313	.333	.134	.356		.370
	(.037)	(.039)	(.036)	(.029)	(.032)	(.029)	(.029)	(.032)	(.039)	(.029)		(.019)
19: IVF	.363	.340	.539	.465	.489	.430	.338	.453	.264	.439	.408	.430
	(.050)	(.055)	(.028)	(.031)	(.030)	(.029)	(.032)	(.038)	(.056)	(.030)	(.029)	(.020)

duced three scales. Only the longest scale, which is the most similar to the Dutch scale, will be considered here. The scale consisted of eight items, measuring tolerance with respect to interpersonal issues. The following items were included in this scale: Do you justify ... adultery (item 6); divorce (item 10); euthanasia (item 11); having casual sex (item 14); prostitution (item 16); experiments on human embryos (item 17); manipulation of food (item 18); and invitro fertilization (item 19). All item pairs had positive $\widehat{H}_{ij}$ values. However, item 16 (prostitution) had an $\widehat{H}_j$ value which was lower than the generally accepted lower bound value .3 (i.e., $\widehat{H}_{16} = .269$; s.e. $= .066$) and was thus removed from the scale. The fact that an item with an H_j value lower than the lower bound c was selected into the scale is an artifact of the method. However, at the moment that the item was selected into the scale, its H_j value with respect to the items already selected at that point was in excess of c. Once an item has been selected, it cannot be deselected anymore (Sijtsma & Molenaar, 2002, pp. 79-80).

A second Mokken scale analysis was performed on the remaining seven items, and Table 3.3 shows the sample values of H_{ij} and H_j, and their asymptotic standard errors. To assess whether the item pair scalability coefficients were greater than zero, 95% confidence intervals were obtained in a similar way to the Dutch sample. Because the value zero was not included in the confidence interval for any of the 21 $\widehat{H}_{ij}$s, all $\widehat{H}_{ij}$s were significantly greater than zero. Also, 95% confidence intervals were created for H_j. For items 6 ($\widehat{H}_6 = .333$; s.e. $= .039$) and 14 ($\widehat{H}_{14} = .345$; s.e. $= .035$), the confidence intervals included the criterion value $c = .3$. So we do not have sufficient evidence that both items satisfy the second property of a Mokken scale, and thus they may be considered for removal from the scale. The sample value for coefficient H was equal to .402 with a standard error of .028. Although the sample value of H suggests that the items are moderately scalable according to Mokken's guidelines, using the standard errors suggests that we can only claim that the items are weakly scalable.

3.5 Discussion

For many sample statistics — for example, correlation coefficients, sample means, and regression parameters — standard errors are vital for the inter-

Table 3.3: *Scalability Coefficients H_{ij} and H_j and Their Standard Errors (Between Brackets) for 7 Items Measuring Tolerance with Respect to Interpersonal Issues for the Georgian Sample.*

	H_{ij}						H_j
	6	10	11	14	17	18	
6: Adultery							.333
							(.039)
10: Divorce	.399						.476
	(.058)						(.031)
11: Euthanasia	.324	.595					.416
	(.054)	(.040)					(.031)
14: Casual Sex	.531	.451	.362				.345
	(.055)	(.058)	(.049)				(.035)
17: Human Embryos	.253	.419	.418	.230			.394
	(.053)	(.058)	(.052)	(.048)			(.038)
18: Manip. Food	.278	.484	.410	.318	.556		.436
	(.059)	(.060)	(.064)	(.061)	(.057)		(.042)
19: IVF	.254	.452	.364	.275	.462	.514	.392
	(.057)	(.039)	(.038)	(.048)	(.044)	(.056)	(.028)

pretation of the size of the effect of the estimated value. This is also true for scalability coefficients, but until recently their standard errors could not be computed. This chapter showed how to derive these standard errors. Although the derivation may be technically difficult, in practice the computation of the standard errors is accomplished by means of the R-package *mokken* (Van der Ark, 2007, 2012), which is available without charge.

In general, it is well-known that standard errors decrease as the sample size N increases (e.g., Tabachnick & Fidell, 2007). However, the standard errors of the scalability coefficients are functions not only of the sample size, but also of the skewness of the item-score distributions. The more skewed the item-score distributions are, the larger the size of the standard errors (Agresti, 2007, p. 110); this is due to estimates of certain coefficients becoming less accurate as the estimated item step proportions get closer to 0 or 1. Consequently, even with a large sample size standard errors can be large. This makes it even more important to consider standard errors when interpreting scalability coefficients.

In our data analysis, we argued that sample values of the scalability coefficients should be significantly greater than the desired criterion, and we investigated each scalability coefficient separately without correction for multiple testing. These two decisions may be open for debate. In statistical hypothesis testing, the null hypothesis usually states the opposite of what one wants to prove (note that this is not the case, for example, in model selection tests in structural equation modelling). Because we want to test whether the item scalability coefficients are greater than .3, the null hypothesis is $H_j \leq .3$. If the burden of proof is reversed, researchers may be tempted to use very small samples (yielding very large confidence intervals), which means that even for low values of H_j and H, the guidelines are met.

When the number of items is large, there will also be a large number of item pair and item scalability coefficients. If for all these H_{ij}s and H_js confidence intervals are constructed simultaneously, the chance of incorrectly rejecting the true null hypothesis (i.e., $H_{ij} \leq 0$; and $H_j \leq c$) is much larger. The probability of obtaining a Type I error will be much larger than it would be when testing one hypothesis at the time. A correction for this multiple hypothesis testing might be used, for example, the Holm-Bonferroni correction (Holm, 1979), which is suited for correlated tests. This results in larger confidence intervals (i.e., 99% or 99.9%), but it may be noted that larger confidence intervals result in a smaller power.

An issue that remains to be solved is that the order of the $2z$ item steps (Equation 3.3) is obtained from the data. In most cases, it is assumed that the ordering of the item steps in the data is identical to the ordering of the item steps in the population. However, when the popularity of two item steps are almost equal in the population, the ordering may be reversed in the sample. This affects the Guttman weights in matrix $\mathbf{A}_3$, because the number of Guttman errors for each item-score pattern depends on the ordering of the item steps. As a result, the reversal may affect the estimates of the scalability coefficients and their standard errors. Investigating the effect of differences in the ordering of item steps between sample and population on the estimates of the scalability coefficients and their standard errors is a topic for future research.

Another topic for future research is to investigate how standard errors af-

fect the automated item selection procedure in Mokken scale analysis. Items are now selected into a scale if all sample values of $H_j \geq c$. But as our examples showed, this may be too liberal as not all sample values of H_j are significantly greater than c.

3.A Derivation of Design Matrices for Item Pair Scalability Coefficients

The $2B \times (B + U + 1)$ design matrix $\mathbf{A}_2$ in Equation 3.15 is used for constructing the expected bivariate frequencies (Equation 3.5). $\mathbf{A}_2$ consists of several submatrices:

$$\mathbf{A}_2 = \begin{pmatrix} \mathbf{I}_B & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} & -\mathbf{1}_B \end{pmatrix}.$$

Matrix $\mathbf{I}_B$ is an identity matrix of order B; multiplying with $\mathbf{I}_B$ leaves the observed bivariate frequencies unchanged. The $B \times U$ submatrix $\mathbf{P}$ is necessary to obtain the B products of observed univariate frequencies (numerator on the right-hand side of Equation 3.5). The first $(z+1)^2$ rows correspond to the first item pair (item 1, item 2); the next $(z+1)^2$ rows correspond to the second item pair (item 1, item 3), and so on; the U columns correspond to the U observed univariate frequencies. Element (p, u) equals 1 if the u-th observed univariate frequency contributes to the p-th product of observed univariate frequencies, and element (p, u) equals 0 otherwise. Vector $-\mathbf{1}_B$ is used for dividing the product of observed univariate frequencies (obtained using matrix $\mathbf{P}$) by N; this results in the expected bivariate frequencies under independence (Equation 3.5). Let $\mathbf{e}_{ij} = (e_{ij}^{00}, e_{ij}^{01}, \ldots, e_{ij}^{zz})^T$ contain the $(z+1)^2$ expected bivariate frequencies pertaining to item i and item j. Substituting $\mathbf{A}_1.\mathbf{n}$ by the right-hand side of Equation 3.17, we find that for the vector of observed frequencies in Equation 3.9 $\exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{n}))$ equals

$$\exp\left(\begin{pmatrix} \mathbf{I}_B & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} & -\mathbf{1}_B \end{pmatrix} \cdot \log\begin{pmatrix} \mathbf{n}_{ab} \\ \mathbf{n}_{ac} \\ \mathbf{n}_{bc} \\ \mathbf{n}_a \\ \mathbf{n}_b \\ \mathbf{n}_c \\ N \end{pmatrix}\right) = \begin{pmatrix} \mathbf{n}_{ab} \\ \mathbf{n}_{ac} \\ \mathbf{n}_{bc} \\ \mathbf{e}_{ab} \\ \mathbf{e}_{ac} \\ \mathbf{e}_{bc} \end{pmatrix} \tag{3.20}$$

The $(2K+1) \times 2B$ design matrix $\mathbf{A}_3$ is used to compute the weighted observed and expected frequencies; it has the following form:

$$\mathbf{A}_3 = \begin{pmatrix} \mathbf{c}_1^T & \mathbf{0} \\ \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{pmatrix}. \tag{3.21}$$

Let $\mathbf{w}_{ij} = (w_{ij}^{00}, w_{ij}^{01}, \ldots, w_{ij}^{zz})^T$ contain the $(z+1)^2$ Guttman weights (Equation 3.4) pertaining to item-pair (i, j), then the $K \times B$ matrix $\mathbf{W}$ is a block-diagonal matrix:

$$\mathbf{W} = \begin{pmatrix} \mathbf{w}_{12}^T & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{w}_{13}^T & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{w}_{14}^T & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{w}_{J-1,J}^T \end{pmatrix}.$$

Vector $\mathbf{c}_1^T$ is a copy of the first row of $\mathbf{W}$; duplicating this row is necessary for constructing scalar 1 in Equation 3.6. Substituting $\exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))$ with the right-hand side of Equation 3.20, we find that for the vector of observed frequencies in Equation 3.9 $\mathbf{A_3} \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))$ equals

$$\begin{pmatrix} \mathbf{c}_1^T & \mathbf{0} \\ \mathbf{W} & \mathbf{0} \\ \mathbf{0} & \mathbf{W} \end{pmatrix} \begin{pmatrix} \mathbf{n}_{ab} \\ \mathbf{n}_{ac} \\ \mathbf{n}_{bc} \\ \mathbf{e}_{ab} \\ \mathbf{e}_{ac} \\ \mathbf{e}_{bc} \end{pmatrix} = \begin{pmatrix} \mathbf{w}_{ab}\mathbf{n}_{ab} \\ \mathbf{w}_{ab}\mathbf{n}_{ab} \\ \mathbf{w}_{ac}\mathbf{n}_{ac} \\ \mathbf{w}_{bc}\mathbf{n}_{bc} \\ \mathbf{w}_{ab}\mathbf{e}_{ab} \\ \mathbf{w}_{ac}\mathbf{e}_{ac} \\ \mathbf{w}_{bc}\mathbf{e}_{bc} \end{pmatrix} = \begin{pmatrix} F_{ab} \\ F_{ab} \\ F_{ac} \\ F_{bc} \\ E_{ab} \\ E_{ac} \\ E_{bc} \end{pmatrix}. \tag{3.22}$$

It may be noted that F_{ij} and E_{ij} were introduced in Equation 3.6.

The $(K+1) \times (2K+1)$ design matrix $\mathbf{A}_4$ is a concatenation of several submatrices,

$$\mathbf{A}_4 = \left(\begin{array}{c|c|c} 1 & -1 & \mathbf{0}_{K-1}^T \quad \mathbf{0}_K^T \\ \hline \mathbf{0}_K & \mathbf{I}_K & -\mathbf{I}_K \end{array} \right). \tag{3.23}$$

Substituting expression $\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))$ with the right-hand side of Equation 3.22, we find that for the vector of observed frequencies in Equa-

tion 3.9 $\exp(\mathbf{A_4}\log(\mathbf{A_3}\exp(\mathbf{A}_2\log(\mathbf{A}_1\mathbf{n}))))$ equals

$$\exp\left(\left(\begin{array}{c|ccc|ccc} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{array}\right)\log\left(\begin{array}{c} F_{ab} \\ F_{ab} \\ F_{ac} \\ F_{bc} \\ E_{ab} \\ E_{ac} \\ E_{bc} \end{array}\right)\right) = \left(\begin{array}{c} 1 \\ F_{ab}/E_{ab} \\ F_{ac}/E_{ac} \\ F_{bc}/E_{bc} \end{array}\right). \tag{3.24}$$

The $K \times (K+1)$ design matrix $\mathbf{A}_5$ is a concatenation of a unit vector of length K, and the negative of an identity matrix of order K; that is,

$$\mathbf{A}_5 = \left(\begin{array}{cc} \mathbf{1}_K & -\mathbf{I}_K \end{array}\right). \tag{3.25}$$

Substituting $\exp(\mathbf{A_4}\log(\mathbf{A_3}\exp(\mathbf{A}_2\log(\mathbf{A}_1\mathbf{n}))))$ by the right-hand side of Equation 3.24, we find that for the vector of observed frequencies in Equation 3.9 $\mathbf{A}_5\exp(\mathbf{A_4}\log(\mathbf{A_3}\exp(\mathbf{A}_2\log(\mathbf{A}_1\mathbf{n}))))$ equals

$$\left(\begin{array}{c|ccc} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{array}\right)\left(\begin{array}{c} 1 \\ F_{ab}/E_{ab} \\ F_{ac}/E_{ac} \\ F_{bc}/E_{bc} \end{array}\right) = \left(\begin{array}{c} 1 - F_{ab}/E_{ab} \\ 1 - F_{ac}/E_{ac} \\ 1 - F_{bc}/E_{bc} \end{array}\right) = \left(\begin{array}{c} H_{ab} \\ H_{ac} \\ H_{bc} \end{array}\right).$$

3.B Derivation of Design Matrices for Item Scalability Coefficients

Matrix $\mathbf{A}_3^\dagger$ can be obtained by premultiplying matrix $\mathbf{A}_3$ (Equation 3.21) by a $(2J+1) \times (2K+1)$ matrix $\mathbf{S}^\dagger$: For $i = 1, 2, \ldots J-1$, let $\mathbf{J}_{i,J}$ be the $J \times (J-i)$ matrix

$$\mathbf{J}_{i,J} = \left(\begin{array}{c} \mathbf{0}_{(i-1)\times(J-i)} \\ \mathbf{1}^T_{1\times(J-i)} \\ \mathbf{I}_{J-i} \end{array}\right),$$

and let $\mathbf{J} = (\mathbf{J}_{1,J}\ \mathbf{J}_{2,J}\ \ldots\ \mathbf{J}_{J-1,J})$; then

$$\mathbf{S}^\dagger = \left(\begin{array}{ccc} 0 & \mathbf{c}_1^T & \mathbf{0} \\ \mathbf{0} & \mathbf{J} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{J} \end{array}\right).$$

Vector $\mathbf{c}_1^T$ is a copy of the first row of $\mathbf{J}$. Matrix $\mathbf{S}^\dagger$ is required in order to add up over the appropriate coefficients F_{ij} and E_{ij} (Equation 3.7). Substituting $\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))$ with the right-hand side of Equation 3.22, we find that for the the vector of observed frequencies in Equation 3.9 $\mathbf{S}^\dagger \mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))$ equals

$$\left(\begin{array}{c|ccc|ccc} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{array}\right) \left(\begin{array}{c} F_{ab} \\ F_{ab} \\ F_{ac} \\ F_{bc} \\ E_{ab} \\ E_{ac} \\ E_{bc} \end{array}\right) = \left(\begin{array}{c} \sum_{i \neq a} F_{ia} \\ \sum_{i \neq a} F_{ia} \\ \sum_{i \neq b} F_{ib} \\ \sum_{i \neq c} F_{ic} \\ \sum_{i \neq a} E_{ia} \\ \sum_{i \neq b} E_{ib} \\ \sum_{i \neq c} E_{ic} \end{array}\right).$$

Design matrices $\mathbf{A}_4^\dagger$ and $\mathbf{A}_5^\dagger$ are very similar to $\mathbf{A}_4$ (Equation 3.23) and $\mathbf{A}_5$ (Equation 3.25), respectively. The only difference is that the sizes of the submatrices are J rather than K.

3.C Derivation of Design Matrices for the Total-Scale Scalability Coefficient

Matrix $\mathbf{A}_3^\ddagger$ can be obtained by premultiplying matrix $\mathbf{A}_3$ (Equation 3.21) by a $3 \times (2K + 1)$ matrix $\mathbf{S}^\ddagger$

$$\mathbf{S}^\ddagger = \begin{pmatrix} 0 & \mathbf{1}_K^T & \mathbf{0}_K^T \\ 0 & \mathbf{1}_K^T & \mathbf{0}_K^T \\ 0 & \mathbf{0}_K^T & \mathbf{1}_K^T \end{pmatrix}.$$

Matrix $\mathbf{S}^\ddagger$ is required in order to add up over the appropriate coefficients F_{ij} and E_{ij} (Equation 3.8). Substituting $\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))$ with the right-hand side of Equation 3.22, we find that for the vector of observed frequencies in Equation 3.9 $\mathbf{S}^\ddagger \mathbf{A_3} \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))$ equals

$$\left(\begin{array}{c|ccc|ccc} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right) \left(\begin{array}{c} F_{ab} \\ F_{ab} \\ F_{ac} \\ F_{bc} \\ E_{ab} \\ E_{ac} \\ E_{bc} \end{array}\right) = \left(\begin{array}{c} \sum \sum_{i \neq j} F_{ij} \\ \sum \sum_{i \neq j} F_{ij} \\ \sum \sum_{i \neq j} E_{ij} \end{array}\right).$$

Using design matrices

$$\mathbf{A}_4^{\ddagger} = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \text{ and } \mathbf{A}_5^{\ddagger} = \begin{pmatrix} 1 & -1 \end{pmatrix}$$

in Equation 3.19 yields coefficient H.

3.D Deriving the Matrix of Partial Derivatives

The Jacobian $\mathbf{G}$ is derived by means of a recursive procedure that requires the design matrices derived in Appendices 3.A, 3.B, and 3.C. First, let $\phi(x)$ be a function that either indicates an exponential ($\phi(x) = \exp(x)$, $\phi'(x) = \exp(x)$), a logarithm ($\phi(x) = \log(x)$, $\phi'(x) = 1/x$), or a translation ($\phi(x) = x + c$, where c is some constant value, $\phi'(x) = 1$). Second, let $\mathbf{f}_0(\mathbf{n}), \mathbf{f}_1(\mathbf{n}), \mathbf{f}_2(\mathbf{n}), \ldots, \mathbf{f}_q(\mathbf{n})$ be a series of $q+1$ functions, in which

$$\begin{aligned} &\mathbf{f}_0(\mathbf{n}) = \mathbf{n}, \\ &\mathbf{f}_i(\mathbf{n}) = \phi[\mathbf{A}_i \mathbf{f}_{i-1}(\mathbf{n})]; \text{ for } i = 1, \ldots, q. \end{aligned} \tag{3.26}$$

The last function in Equation 3.26 is

$$\mathbf{f}_q(\mathbf{n}) = \mathbf{g}(\mathbf{n})$$

For example, for scalability coefficient H_{ij} in Equation 3.15, $\mathbf{f}_0(\mathbf{n}) = \mathbf{n}$, $\mathbf{f}_1(\mathbf{n}) = \log(\mathbf{A}_1 \mathbf{f}_0(\mathbf{n})) = \log(\mathbf{A}_1 \mathbf{n})$, $\mathbf{f}_2(\mathbf{n}) = \exp(\mathbf{A}_2 \mathbf{f}_1(\mathbf{n})) = \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{n}))$, and so forth until $\mathbf{f}_5(\mathbf{n}) = \mathbf{A}_5 \mathbf{f}_4(\mathbf{n}) = \mathbf{g}(\mathbf{n})$. Third, the following recursive relationship can be derived for the partial derivatives of the functions $\mathbf{f}_i(\mathbf{n})$:

$$\frac{\partial \mathbf{f}_0(\mathbf{n})}{\partial \mathbf{n}^T} = \mathbf{I},$$

and

$$\frac{\partial \mathbf{f}_i(\mathbf{n})}{\partial \mathbf{n}^T} = \mathbf{D}\left[\boldsymbol{\phi}'(\mathbf{A}_i \mathbf{f}_{i-1})\right] \mathbf{A}_i \frac{\partial \mathbf{f}_{i-1}(\mathbf{n})}{\partial \mathbf{n}^T}; \text{ for } i = 1, \ldots q. \tag{3.27}$$

It may be noted that if ϕ indicates an exponential, then Equation 3.27 equals

$$\frac{\partial \mathbf{f}_i(\mathbf{n})}{\partial \mathbf{n}^T} = \mathbf{D}\left[\exp(\mathbf{A}_i \mathbf{f}_{i-1})\right] \mathbf{A}_i \frac{\partial \mathbf{f}_{i-1}(\mathbf{n})}{\partial \mathbf{n}^T};$$

if ϕ indicates a logarithm, then Equation 3.27 equals

$$\frac{\partial \mathbf{f}_i(\mathbf{n})}{\partial \mathbf{n}^T} = \mathbf{D}^{-1}(\mathbf{A}_i \mathbf{f}_{i-1}) \mathbf{A}_i \frac{\partial \mathbf{f}_{i-1}(\mathbf{n})}{\partial \mathbf{n}^T};$$

and if ϕ indicates a translation, then Equation 3.27 equals

$$\frac{\partial \mathbf{f}_i(\mathbf{n})}{\partial \mathbf{n}^T} = \mathbf{A}_i \frac{\partial \mathbf{f}_{i-1}(\mathbf{n})}{\partial \mathbf{n}^T}.$$

Fourth, the Jacobian can be obtained as

$$\mathbf{G} = \frac{\partial \mathbf{f}_q(\mathbf{n})}{\partial \mathbf{n}^T}.$$

For example, to obtain the Jacobian of $\mathbf{H}_{ij}$ in Equation 3.15, Equation 3.27 is applied recursively for $i = 1, 2, 3, 4, 5$.

The recursive procedure in Equation 3.27 for $i = 1, 2, 3$ can be avoided by computing $\mathbf{f}_3$ and $\partial \mathbf{f}_3(\mathbf{n}^*)/\partial \mathbf{n}^{*T}$ directly from the data. This has the advantage that the first three design matrices do not need to be computed separately. In the recursive procedure described above, for $i = 3$ and for the reduced vector $\mathbf{n}^*$, Equation 3.27 equals

$$\begin{aligned} \frac{\partial \mathbf{f}_3(\mathbf{n}^*)}{\partial \mathbf{n}^{*T}} &= \mathbf{D}\left[\boldsymbol{\phi}'(\mathbf{A}_3\mathbf{f}_2)\right]\mathbf{A}_3 \frac{\partial \mathbf{f}_2(\mathbf{n}^*)}{\partial \mathbf{n}^{*T}} \\ &= \mathbf{A}_3\mathbf{D}\left[\exp(\mathbf{A}_2 \log(\mathbf{A}_1^*\mathbf{n}^*))\right]\mathbf{A}_2\mathbf{D}^{-1}\left[\mathbf{A}_1^*\mathbf{n}^*\right]\mathbf{A}_1^*. \end{aligned} \tag{3.28}$$

Let $\mathbf{M}^*$ be a $B \times L^*$ matrix relating the B bivariate frequencies to the L^* observed response patterns. Suppose that the b-th row of $\mathbf{M}^*$ pertains to bivariate frequency n_{ij}^{xy}, then element (b, l) equals $e_{ij}^{xy}/n_i^x + e_{ij}^{xy}/n_j^y - 1$ if in the l-th response pattern the score on item i equals x and the score on item j equals y; element (b, l) equals $e_{ij}^{xy}/n_i^x - 1$ if in the l-th response pattern the score on item i equals x and the score on item j does not equal y; element (b, l) equals $e_{ij}^{xy}/n_j^y - 1$ if in the l-th response pattern the score on item i does not equal x and the score on item j equals y; and element (b, l) equals -1 if in the l-th response pattern the score on item i does not equal x and the score on item j does not equal y. Let $\mathbf{B}^*$ (of order $B \times L^*$) be the reduced version of matrix $\mathbf{B}$ introduced in Equation 3.16, and let $\mathbf{W}$ be the $K \times B$ matrix of Guttman weights (see Appendix 3.A, Equation 3.21). Tedious yet straightforward algebra shows that Equation 3.28 is equal to

$$\frac{\partial \mathbf{f}_3(\mathbf{n}^*)}{\partial \mathbf{n}^{*T}} = \begin{pmatrix} \mathbf{c}_1^T \\ \mathbf{W}\mathbf{B}^* \\ \mathbf{W}\mathbf{M}^* \end{pmatrix},$$

where $\mathbf{c}_1^T$ is a copy of the first row of $\mathbf{WB}^*$. The proof can be obtained from the author of this dissertation. For the subsequent steps in the recursive procedure described in this Appendix, $\mathbf{f}_3 = \mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1^* \mathbf{n}))$ equals $(F_{12}, F_{12}, F_{13}, \ldots, F_{J-1,J}, E_{12}, E_{13}, \ldots, E_{J-1,J})$ (cf. Equation 3.22 in Appendix 3.A) which can be computed directly from the data. It may be noted that $\mathbf{c}_1^T$ yields a duplication of F_{12}.

3.E Data and R Code of Examples

The empirical data used in the two real-data examples were collected in the 2008 wave of the European Values Study (EVS, 2011). From these data, two data sets have been made available on the following website: https://sites.google.com/site/rekuijpers/data-sets. Data set `EVS2008.NL` contains the scores on the 12 tolerance items pertaining to the largest Mokken scale for the Dutch sample, and `EVS2008.GE` contains the scores on the 7 tolerance items pertaining to the largest Mokken scale for the Georgian sample. In both data sets the items have been recoded from ten into three categories, and cases with missings have been deleted. The R code installs the R package *mokken*, reads the data, and computes the scalability coefficients and their standard errors. Following R conventions, `R>` indicates the R prompt.

```
R> # Install mokken package if necessary.
R> if(is.na(packageDescription("mokken")[[1]]))
   install.packages("mokken")
R> library(mokken)

R> # Read data
R> EVS2008.NL <- read.table(file="EVS08NL.txt")
R> EVS2008.GE <- read.table(file="EVS08GE.txt")

R> # Compute scalability coefficients and standard errors.
R> coefH(EVS2008.NL)
R> coefH(EVS2008.GE)
```

Chapter 4

Bias in Estimates and Standard Errors of Mokken's Scalability Coefficients

Abstract In Mokken scale analysis, there are three types of scalability coefficients: (1) for pairs of items, (2) for items, and (3) for the entire scale. Recently, standard errors for the scalability coefficients were derived by means of marginal models. Both the estimates and the standard errors of the scalability coefficients assume that the ordering of the item steps in the sample is identical to the ordering of the item steps in the population. Due to sampling error, the sample ordering may be incorrect and, as a result, the estimates and standard errors may be biased. In two simulation studies, we investigated the bias of the estimates and the standard errors of the scalability coefficients and the coverage of the 95% confidence intervals. In addition to the most important design factor, distance between item steps, we included sample size, number of items, and number of answer categories. Bias for the standard errors was negligible in all cases. Bias of the estimates was largest for identical items steps, especially for small sample sizes. Furthermore, bias of the estimates decreased as the number of answer categories increased. Coverage of the 95% confidence intervals was close to .950 for all cases, only for small sample sizes the coverage was slightly poorer. Coverage of the confidence

This chapter is submitted for publication as Kuijpers, R.E., Van der Ark, L.A., Croon, M.A., & Sijtsma, K. Bias in estimates and standard errors of Mokken's scalability coefficients.

intervals became poorer as numbers of items increased, particularly when items were dichotomous.

4.1 Introduction

Mokken scale analysis (Mokken, 1971; also see, e.g., Sijtsma & Molenaar, 2002) is an important statistical tool for the construction of tests and questionnaires in social science research. Among other model assessment methods, Mokken scale analysis involves an item selection algorithm that partitions a set of items into one or more unidimensional scales. The analysis is based on a nonparametric item response theory (NIRT) model, which allows all items with nondecreasing item response functions (IRFs) to be included in a scale (Sijtsma & Molenaar, 2002). Mokken scale analysis is extensively used in test construction in various research areas. Recent examples of its use in psychology include tests assessing psychological distress and well-being (Watson, Wang, Thompson, & Meijer, 2014), depression and anxiety (Bech, Bille, Moller, Hellström, & Ostergaard, 2014; Bedford, Watson, Henry, Crawford, & Deary, 2011), disability in activities of daily living (Kingston et al., 2012), learning disability (Murray & McKenzie, 2013), and sexual sadism (Nitschke, Osterheider, & Mokros, 2009).

In scale construction, Mokken scale analysis uses three types of scalability coefficients, as criteria for item set partitioning and as diagnostics for the strength of the scales. These are item pair coefficient H_{ij}, assessing the scalability of item pair (i, j); item coefficient H_j, assessing the scalability of item j; and total-scale coefficient H, assessing the scalability of the entire scale. In Mokken scale analysis, items are defined to form a scale if all item pair scalability coefficients H_{ij} are positive, and if the item scalability coefficients H_j are at least equal to a positive lower bound c which, based on experience, is chosen to be .3.

In order to compute a scalability coefficient, one needs the ordering of the item steps (Molenaar, 1991; to be discussed in detail below). The item step ordering is obtained from the sample. For estimating the scalability coefficients, it is assumed that the sample ordering of the item steps is identical to the population ordering. Due to sampling fluctuation, the estimated ordering of the item steps may be different from the ordering of the item steps in the

population. Consequently, the estimated scalability coefficients may use an incorrect item step ordering, which causes the estimates to be biased. This means that the scalability coefficients either underestimate or overestimate their parameter values. For dichotomous items, based on statistical reasoning involving all the 2×2 tables, Sijtsma and Molenaar (2002, p. 56) proposed that this bias is almost negligible for $N > 200$ when incidental pairs of item steps are close together ($< .02$) and for $N > 400$ when many item steps are close together. From their discussion it is clear that additional research may be needed to support the rules of thumb. Recently, Kuijpers et al. (2013b) analytically derived standard errors for each type of scalability coefficient. The standard errors are also based on the assumption that the ordering of the item steps in the sample and the population are the same and, as a consequence, the standard errors may also be biased. Biased standard errors are either too small or too large, and produce confidence intervals having an incorrect coverage. The magnitude of the bias is unknown, and therefore needs to be investigated.

Using two simulation studies, we investigated the effect of differences in the item step ordering between the sample and population on the estimates of the scalability coefficients and the standard errors. We assessed the bias of the estimates and the standard errors, and the coverage of the confidence intervals under several conditions. The most important independent factor in the experimental design was the distance between the item steps in the population; a smaller distance increases the probability that the item step ordering in the sample and population are different. Other independent factors were sample size, number of items, and number of answer categories.

This study is organized as follows. First, we discuss Mokken scale analysis and the scalability coefficients. Second, we briefly explain the computation of the standard errors by means of the marginal modelling approach. Third, we discuss the design of the two simulation studies. Fourth, we present the results of the first simulation study. Fifth, we discuss the setup of the second simulation study, and the results. Finally, we discuss the general results and we give recommendations about the use of the standard errors in different situations.

4.2 Mokken Scale Analysis

4.2.1 The Monotone Homogeneity Model

Mokken scale analysis is based on the monotone homogeneity model (Mokken, 1971, Chapter 4; Sijtsma & Molenaar, 2002, pp. 22-23), which is a NIRT model for measuring respondents on an ordinal scale. Let θ denote a latent variable that underlies the performance on each item in the test. For a set of J items each with $z+1$ ordered answer categories $x = 0, 1, \ldots, z$, $P(X_j = x|\theta)$ denotes the probability of having a score x on item j, and $P(X_1 = x_1, X_2 = x_2, \ldots, X_J = x_J|\theta)$ denotes the probability of a particular item score pattern on all J items. Furthermore, $X_+ = \sum_j^J X_j$ denotes the total score on the J items. The monotone homogeneity model is based on the following assumptions:

Unidimensionality: All J items measure the same latent variable θ, hence, θ is unidimensional;

Local independence: The item scores are independent given the latent variable θ; that is, $P(X_1 = x_1, X_2 = x_2, \ldots, X_J = x_J|\theta) = \prod_{j=1}^{J} P(X_j = x_j|\theta)$.

Monotonicity: As θ increases, the probability of having a score of at least x on item j is nondecreasing; that is, $P(X_j \geq x|\theta)$ is a nondecreasing function of θ.

For dichotomous items, the monotone homogeneity model implies the stochastic ordering of θ by means of total sum score X_+. For polytomous items, the monotone homogeneity model implies a weaker stochastic ordering property; for details, see Van der Ark and Bergsma (2010). If the model fits the data well, the stochastic ordering properties can be used for ordering respondents on latent variable θ by means of total score X_+.

In Mokken scale analysis, several manifest properties of the monotone homogeneity model are investigated so as to establish fit of the model to the data and find support for the use of X_+ as an ordinal estimator of θ. For example, the model implies that the covariances between all item pairs are nonnegative. For a set of items meant to constitute a scale, this property

can be investigated by evaluating whether the sample values of all item pair scalability coefficients are nonnegative.

4.2.2 Scalability Coefficients

Item Steps and Weighted Guttman Errors

The three scalability coefficients used in Mokken scale analysis are based on the item step ordering in each pair of items and the corresponding weighted Guttman errors (Molenaar, 1991; also see Kuijpers et al., 2013b). A single item j having $z+1$ ordered answer categories has z ordered item steps: $X_j \geq 1$, $X_j \geq 2$, ..., $X_j \geq z$. It may be noted that this ordering is the same for each respondent. Obtaining a score x on an item j can be regarded as passing item steps $X_j \geq 1, \ldots, X_j \geq x$ and failing item steps $X_j \geq x+1, \ldots, X_j \geq z$. Consider dichotomous variable Y_j^x, which takes on a value 1 if the respondent has passed an item step ($X_j \geq x$) and a value 0 if the respondent failed an item step ($X_j < x$); then, $X_j = \sum_{u=1}^{x} Y_j^u$. The item steps are ordered by their popularity, which is the probability that a randomly chosen respondent passes the item step: $P(X_j \geq x)$. It may be noted that $X_j \geq 0$ usually is not considered to be an item step, because by definition $P(X_j \geq 0) = 1$. If in a particular item a less popular item step is passed, by definition the more popular step is also passed.

For item pair i and j, the item steps are $X_i \geq 1, \ldots, X_i \geq z$ and $X_j \geq 1, \ldots, X_j \geq z$. The ordering of the $2z$ item steps may not be the same for each respondent, and some individuals may pass a less popular item step and fail a more popular item step. According to the Guttman (1950) model, this incidence is referred to as a Guttman error (Molenaar, 1991). Table 4.1 shows an example of the joint probabilities of having a score x on item a and a score y on item b; that is, $P(X_a = x, X_b = y)$ with $x, y = 0, 1, 2, 3$. The marginal probabilities are defined by $P(X_a = x)$ and $P(X_b = y)$, and the cumulative probabilities by $P(X_a \geq x)$ and $P(X_b \geq y)$. For this example, the cumulative probabilities order the item steps by descending popularity as

$$X_b \geq 1, X_a \geq 1, X_b \geq 2, X_b \geq 3, X_a \geq 2, X_a \geq 3. \tag{4.1}$$

Let index h enumerate the number of most popular item steps passed. Item-

Table 4.1: *Cross-Tabulation of Probability of Obtaining Particular Item-Score Pattern; Guttman Weights Are Between Parentheses.*

	X_b					
X_a	0	1	2	3	$P(X_a = x)$	$P(X_a \geq x)$
0	**.044** (0)	**.013** (0)	.019 (1)	.025 (2)	.101	1.000
1	.023 (1)	**.060** (0)	**.106** (0)	**.267** (0)	.456	.899
2	.011 (4)	.028 (2)	.193 (1)	**.145** (0)	.377	.443
3	.002 (7)	.012 (4)	.042 (2)	**.010** (0)	.066	.066
$P(X_b = y)$	.080	.113	.360	.447	1.000	
$P(X_b \geq y)$	1.000	.920	.807	.447		

Note: Probabilities of item-score patterns that are in agreement with the Guttman model are printed in boldface.

score patterns (0,0), (0,1), (1,1), (1,2), (1,3), (2,3), and (3,3) (see Table 4.1, corresponding probabilities are printed in boldface) are consistent with the Guttman (1950) model because the h most popular item steps in Equation 4.1 are passed, and the remaining $2z - h$ less popular steps are not passed. The remaining item-score patterns are inconsistent with the Guttman model, and to arrive at any of these patterns one or more Guttman errors are made. For example, to obtain item-score pattern (0,3), the more popular item step $X_a \geq 1$ is failed, whereas the less popular item steps $X_b \geq 2$ and $X_b \geq 3$ are passed.

Molenaar (1991) proposed weighing the sample frequencies of the Guttman errors (in Table 4.1, weights are shown between parentheses) depending on the degree to which the item step ordering was violated according to the Guttman model. The weight for a particular item-score pattern ($X_i = x, X_j = y$), denoted w_{ij}^{xy}, is equal to the number of item step pairs for which the less popular step is passed and the more popular step is failed. Ligtvoet et al. (2010), Zijlstra, Van der Ark, and Sijtsma (2011), and Kuijpers et al. (2013b) discussed the computation of these weights. Because the weights play a crucial role in the potential bias in the scalability coefficients and the standard errors, the computation is reiterated here. Consider indicator vector $\mathbf{q}_{ij}^{xy} = (q_{ij(1)}^{xy}, q_{ij(2)}^{xy}, \ldots, q_{ij(2z)}^{xy})$, whose elements correspond to the $2z$ ordered item steps of the item pair (i, j) and take on a value 1 if an item step

has been passed in order to obtain item-score pattern $(X_i = x, X_j = y)$, and a value 0 otherwise. The $2z$ item steps are ordered by descending popularity, as in Equation 4.1. Then, weight w_{ij}^{xy} equals

$$w_{ij}^{xy} = \sum_{u=2}^{2z} q_{ij(u)}^{xy} \left(\sum_{v=1}^{u-1} |1 - q_{ij(v)}^{xy}| \right). \tag{4.2}$$

For each pair of 0s and 1s, Equation 4.2 counts how often a score 0 precedes a score 1 in vector $\mathbf{q}_{ij}^{xy}$. For example, for item-score pattern (1,2) the first three item steps in Equation 4.1 are passed. These are the three most popular steps, implying $\mathbf{q}_{ab}^{12} = (1, 1, 1, 0, 0, 0)$, and because 0 scores do not precede 1 scores, weight $w_{ab}^{12} = 0$. For item-score pattern (0,3), item steps $X_b \geq 1$, $X_b \geq 2$ and $X_b \geq 3$ are passed, so that vector $\mathbf{q}_{ab}^{03} = (1, 0, 1, 1, 0, 0)$. In this case, a 0 score precedes a 1 twice, and thus weight $w_{ab}^{03} = 2$.

Due to sampling fluctuation, the estimated item step ordering may differ across samples. When the item step ordering changes, the weights of the Guttman errors change. As an example, we drew two random samples of 200 observations from the population values in Table 4.1. Table 4.2 shows the joint frequencies for the two samples. In the first sample (upper panel of Table 4.2), the estimated item step ordering is identical to the population item step ordering in Table 4.1. The weights are determined by the estimated item step ordering. Since the estimated ordering is identical to the population ordering the weights of the first sample equal the weights in the population. In the second sample (lower panel of Table 4.2) the estimated item step ordering and the corresponding weights are different from the population values. Using weights different from the population weights may result in biased estimates and standard errors of the scalability coefficients. Molenaar (1991) showed that when two item steps have equal popularities, the scalability coefficients have the same value irrespective of the sample ordering of the two item step popularities.

Scalability Coefficients and Their Standard Errors

For item pair (i, j), scalability coefficient H_{ij} expresses the strength of the association between items i and j corrected for the marginal distributions of their item scores (Van der Ark, Croon, & Sijtsma, 2008a, 2008b), which can

Table 4.2: *Two Samples ($N = 200$) Drawn From the Distribution in Table 4.1. In Sample 1 (Upper Panel), the Estimated Item Step Ordering is Identical to the Item Step Ordering in the Population, Whereas in Sample 2 (Lower Panel), the Estimated Item Step Ordering is Different From Ordering in the Population.*

	X_b					
X_a	0	1	2	3	Freq.	$\widehat{P}(X_a \geq x)$
0	**13** (0)	**1** (0)	2 (1)	4 (2)	20	1.000
1	2 (1)	**10** (0)	**20** (0)	**64** (0)	96	.900
2	2 (4)	2 (2)	40 (1)	**30** (0)	74	.420
3	0 (7)	3 (4)	6 (2)	**1** (0)	10	.050
Freq.	17	16	68	99	200	
$\widehat{P}(X_b \geq y)$	1.000	.915	.835	.495		

	X_b					
X_a	0	1	2	3	Freq.	$\widehat{P}(X_a \geq x)$
0	**8** (0)	**1** (0)	6 (1)	4 (3)	19	1.000
1	6 (1)	**12** (0)	**24** (0)	51 (1)	93	.905
2	3 (3)	7 (1)	**44** (0)	**26** (0)	80	.440
3	0 (6)	2 (3)	5 (1)	**1** (0)	8	.040
Freq.	17	22	79	82	200	
$\widehat{P}(X_b \geq y)$	1.000	.915	.805	.410		

be written in terms of Guttman errors. The H_{ij} coefficient compares the sum of weighted observed Guttman errors for item pair i and j, denoted by F_{ij}, to the sum of weighted Guttman errors expected under marginal independence, denoted by E_{ij}, and is defined as

$$H_{ij} = 1 - \frac{F_{ij}}{E_{ij}}. \tag{4.3}$$

Under the monotone homogeneity model, all H_{ij}'s should be greater than zero (Mokken, 1971, pp. 148-153; Sijtsma & Molenaar, 2002, p. 59).

Item scalability coefficient H_j is used for expressing the strength of the association between item j and the other items in a test (Sijtsma & Molenaar, 2002, p. 36), and combines the information from the $J - 1$ H_{ij}s ($i \neq j$) in which item j is involved. Like H_{ij}, coefficient H_j can be written in terms of Guttman errors. Coefficient H_j compares the sum of weighted observed

Guttman errors to the sum of weighted expected Guttman errors for an individual item, so that

$$H_j = 1 - \frac{\sum_{i \neq j} F_{ij}}{\sum_{i \neq j} E_{ij}}. \tag{4.4}$$

The monotone homogeneity model implies that $0 \leq H_j \leq 1$. For practical purposes, Mokken (1971, p. 184) proposed that in a scale all H_js should be greater than a positive lower bound c, which may be chosen by the researcher but by default is equal to .3 (Sijtsma & Molenaar, 2002, p. 60). This default value is a rule of thumb but it is not entirely arbitrary because items with $H_j < .3$ contribute little to a reliable person ordering, and thus can be best left out of the scale (Sijtsma & Molenaar, 2002, p. 36).

Total-scale scalability coefficient H expresses the degree to which respondents can be ordered by means of a complete set of items (Sijtsma & Molenaar, 2002, p. 39), and is a weighted average of the J coefficients H_j (Mokken & Lewis, 1982). Coefficient H compares the sum of weighted observed Guttman errors to the sum of weighted expected Guttman errors for all J items in a scale, and is defined as

$$H = 1 - \frac{\sum \sum_{i \neq j} F_{ij}}{\sum \sum_{i \neq j} E_{ij}}. \tag{4.5}$$

Mokken (1971, p. 185) proposed that H should be at least equal to .3; values below .3 indicate that the items together do not define a scale. Furthermore, he defined the strength of a scale to be weak if $.3 \leq H < .4$, moderate if $.4 \leq H < .5$, and strong if $H \geq .5$.

In the absence of Guttman errors, the scalability coefficients are equal to 1, and their values decrease as the number of Guttman errors increase. Molenaar (1991) and Sijtsma and Molenaar (2002, p. 57) showed that the scalability coefficients can be written as weighted sums of normed inter-item covariances, where the covariance is normed by dividing it by the maximum possible covariance given the marginal item-score distributions.

Biased H_{ij}, H_j, and H coefficients may influence the composition of a Mokken scale. When H_{ij} or H_j is underestimated, an item might incorrectly be left out of a Mokken scale and when they are overestimated, weakly discriminating items may incorrectly be included in the Mokken scale. A biased

H provides an incorrect assessment of the strength of a scale. Hence, biased estimates and standard errors of the scalability coefficients may produce incorrect conclusions about the inclusion or the exclusion of items in a Mokken scale.

Kuijpers et al. (2013b) used a two-step method based on categorical marginal models to derive asymptotic standard errors for each of the three scalability coefficients. First, data were collected in a frequency vector $\mathbf{n}$, in which the number of elements is equal to the number of item-score patterns in the data. Under the nonrestrictive assumption that vector $\mathbf{n}$ follows a multinomial distribution, the variance-covariance matrix of $\mathbf{n}$, denoted $\mathbf{V_n}$, is well known (e.g., Agresti, 2013). Second, each of the three scalability coefficients was written as a vector function of $\mathbf{n}$, denoted $\mathbf{g}(\mathbf{n})$. Let $\mathbf{G}(\mathbf{n})$ be the matrix of first partial derivatives of $\mathbf{g}(\mathbf{n})$ to $\mathbf{n}$, then according to the delta method the variance-covariance matrix of the scalability coefficients, denoted $\mathbf{V_{g(n)}}$, is approximated by $\mathbf{G(n)V_nG(n)}^T$. The standard errors of the scalability coefficients are obtained by taking the square root of the diagonal elements of $\mathbf{G(n)V_nG(n)}^T$. The derivation of $\mathbf{g}(\mathbf{n})$ and $\mathbf{G}(\mathbf{n})$ is cumbersome; for more details, see Kuijpers et al. (2013b).

4.3 Simulation Study 1

4.3.1 Method

Simulation Model

We simulated data using the graded response model (Samejima, 1969, 1972). This model is a parametric version and hence a special case of the monotone homogeneity model (Hemker, Sijtsma, Molenaar, & Junker, 1996). The model also is a generalization to polytomous items of the 2-parameter logistic model for dichotomous items. The graded response model describes response probabilities to each item j (with scores $x = 0, 1, \ldots, z$) by means of a logistic function with a slope parameter α_j and z location parameters δ_{jx}. Note that for one item, the location parameters are ordered such that $\delta_{jx} < \delta_{j,x+1}$. The probability of a score of at least x on item j equals

$$P(X_j \geq x|\theta) = \frac{\exp[\alpha_j(\theta - \delta_{jx})]}{1 + \exp[\alpha_j(\theta - \delta_{jx})]}. \tag{4.6}$$

Again, by definition $P(X_j \geq 0|\theta) = 1$.

In the simulation study, the slope parameter α_j was chosen to equal 1.5 for all items in all design cells, so that in combination with $\theta \sim N(0, 1)$ and suitable choices for the location parameters the population values of scalability coefficient H had an acceptable value of $c \geq .3$. The values of the location parameters δ_{jx} (Table 4.3) varied across design cells depending on the level of the independent factor 'Distance between item steps' (see below). For each sample, N θ-values were randomly drawn. For each set of θ values, a data set was generated using Equation 4.6 in which the δ_{jx} values (Table 4.3) were inserted.

Design of the Study

In the simulation study, various factors were varied as follows.

Number of items (J). The number of items was equal to either $J = 2$ or $J = 3$. The number of items was small so as to keep the simulation study manageable. Furthermore, because coefficient H is a weighted mean of the pairwise computed $\frac{1}{2}J(J-1)$ H_{ij} coefficients, we expected the bias in the estimates and the standard errors of the scalability

Table 4.3: *Location Parameters for Item Step Orderings*

$z+1$	J	Identical		Close By		Far Away		Extreme	
		δ_j		δ_j		δ_j		δ_j	
2	2	0.000		-0.113		-0.227		-0.343	
		0.000		0.113		0.227		0.343	
	3	0.000		-0.227		-0.460		-0.706	
		0.000		0.000		0.000		0.000	
		0.000		0.227		0.460		0.706	
		δ_{j1}	δ_{j2}	δ_{j1}	δ_{j2}	δ_{j1}	δ_{j2}	δ_{j1}	δ_{j2}
3	2	-0.250	0.250	-0.343	0.113	-0.706	0.227	-1.119	0.343
		-0.250	0.250	-0.113	0.343	-0.227	0.706	-0.343	1.119
	3	-0.250	0.250	-0.581	0.113	-1.278	0.227	-2.563	0.343
		-0.250	0.250	-0.343	0.343	-0.706	0.706	-1.119	1.119
		-0.250	0.250	-0.113	0.581	-0.227	1.278	-0.343	2.563

coefficients and the coverage of the 95% confidence intervals to stay equal as the number of items increases.

Number of answer categories ($z+1$). The items were either dichotomous ($z+1=2$) or polytomous ($z+1=3$). Polytomous items have more item steps than dichotomous items, hence the probability increases that the item step ordering in the sample differs from the ordering in the population. Compared to dichotomous items, for polytomous items we expected more bias in the estimates of the scalability coefficients and their standard errors, and a poorer coverage of the 95% confidence interval.

Sample size (N). The sample size was either small ($N=50$), medium ($N=200$), large ($N=500$), or very large ($N=1500$). As sample sizes become smaller, an extra respondent in an error cell has more influence on the item step ordering. Thus, we expected that as the sample size decreases, the bias of the estimates and the standard errors increases, and we expected the coverage of the 95% confidence interval to deteriorate.

Distance between item steps. The greater the distance between two adjacent item steps, the higher the probability that the sample item step ordering equals the population item step ordering. Distance between item steps had four levels, labeled Identical, Close By, Far Away, and Extreme. The distances between the item steps were varied by manipulating the location parameters δ_{jx} of the graded response model. The ordering of the item steps was fixed to $P(X_1 \geq 1) > P(X_2 \geq 1) > \ldots > P(X_J \geq 1) > P(X_1 \geq 2) > \ldots > P(X_J \geq 2) > \ldots > P(X_1 \geq z) > \ldots > P(X_J \geq z)$. For this ordering, the distance between two consecutive item step probabilities is denoted by Δ. Distance Δ equaled 0, .06, .12, and .18 for the levels Identical, Close By, Far Away, and Extreme, respectively. Table 4.4 shows the resulting cumulative item step probabilities. Once the item step probabilities were fixed, we determined the corresponding location parameters δ_{jx} for the graded response model

in Equation 4.6, such that

$$P(X_j \geq x) = \int P(X_j \geq x|\theta)dG(\theta)$$

equaled the desired values in Table 4.4 for $G(\theta)$, which denotes the cumulative distribution function of θ. Because a smaller Δ value produces a smaller distance between population item step popularities, more reversals of the item step ordering may occur in the sample. Consequently, we expected more bias in the estimates and the standard errors of the scalability coefficients and a poorer coverage of the 95% confidence interval.

The outcome variables of the simulation study were the bias of the estimates of the scalability coefficient H, the bias of the standard errors of H, and the coverage of the 95% confidence interval. The number of replications for each design cell was $Q = 10,000$.

Bias of the estimates (*bias*). Let $\widehat{H}_q$ denote the sample value of coefficient H computed for the qth replication ($q = 1, \ldots, Q$). Let H denote

Table 4.4: *Theoretical Cumulative Item Step Probabilities for Item Step Orderings. Distances are Equal Between Successive Item Steps.*

$z+1$	J	Identical		Close By		Far Away		Extreme	
		π_j		π_j		π_j		π_j	
2	2	.500		.530		.560		.590	
		.500		.470		.440		.410	
	3	.500		.560		.620		.680	
		.500		.500		.500		.500	
		.500		.440		.380		.320	
		π_{j1}	π_{j2}	π_{j1}	π_{j2}	π_{j1}	π_{j2}	π_{j1}	π_{j2}
3	2	.566	.434	.590	.470	.680	.440	.770	.410
		.566	.434	.530	.410	.560	.320	.590	.230
	3	.566	.434	.650	.470	.800	.440	.950	.410
		.566	.434	.590	.410	.680	.320	.770	.230
		.566	.434	.530	.350	.560	.200	.590	.050

Note: The cumulative item step probability $P(X_j \geq 1)$ is in the Table denoted by π_{j1}, the cumulative item step probability $P(X_j \geq 2)$ is denoted by π_{j2}.

the population value of the scalability coefficient, which was obtained by means of linear programming. Then, the bias based on Q replications was computed as

$$bias = \frac{1}{Q}\sum_{q=1}^{Q}(\widehat{H}_q - H). \tag{4.7}$$

Bias of the standard errors (*bias.se*). To estimate the bias of the standard errors we first computed the standard deviation of the estimates of H, denoted $sd(\widehat{H})$, across the Q replications. Let $\bar{H} = \frac{1}{Q-1}\sum_{q=1}^{Q}\widehat{H}_q$, then

$$sd(\widehat{H}) = \sqrt{\frac{1}{Q-1}\sum_{q=1}^{Q}(\widehat{H}_q - \bar{H})^2}. \tag{4.8}$$

Standard deviation $sd(\widehat{H})$ estimates the variability of scalability coefficients across replications, and serves as a gold standard for the standard error. Here, let $se(\widehat{H}_q)$ denote the estimated standard error of the qth estimate of H. Then, the bias of the standard errors equals

$$bias.se = \frac{1}{Q}\sum_{q=1}^{Q}(se(\widehat{H}_q) - sd(\widehat{H})). \tag{4.9}$$

Coverage of the 95% confidence interval To investigate the coverage of the 95% confidence interval we first constructed a confidence interval for each qth replication using $\widehat{H}_q \pm 1,96 * se(\widehat{H}_q)$. Then, the 95% coverage was defined by the proportion of replications for which the 95% confidence interval contains the population value of H.

The population values for coefficient H varied across design cells, and are given in Table 4.5. It may be noted that the population values are unaffected by sample size. The simulation study was programmed in R (R Core Team, 2014), using the R-package *mokken* (Van der Ark, 2007, 2012) to compute the estimates and standard errors of coefficient H for each sample across the 10,000 replications.

Table 4.5: *Population Values for Coefficient H for All Item Step Orderings.*

$z+1$	J	Identical	Close By	Far Away	Extreme
2	2	.293	.329	.366	.404
	3	.293	.340	.386	.431
3	2	.327	.344	.388	.425
	3	.327	.363	.415	.451

4.3.2 Results

The bias of the estimates of scalability coefficient H was less than .05 in all conditions (Figure 4.1). For the conditions with identical item steps, the bias of H was slightly larger for both $J = 2$ (upper panel) and $J = 3$ (lower panel) compared to the other conditions. As expected, an increase of the number of items did not affect the bias in H. For all four different distances between item steps, Figure 4.1 shows that the bias in H decreased as sample size increased; for conditions involving close by or identical item steps the bias was considerably larger for $N = 50$ than for the other sample sizes. Furthermore, inconsistent with our expectation that bias increases as number of answer categories increases, the bias in H was larger for two answer categories than for three.

Table 4.6 shows the bias of the standard errors of coefficient H and the coverage of the 95% confidence intervals for the conditions having identical item steps; these conditions showed the largest bias and the poorest coverage. The results showed that the bias of the standard errors of H was 0 or close to 0 in all design cells, whereas we had expected bias of the standard errors to increase as the number of answer categories increased and sample size decreased.

Coverage of the 95% confidence intervals was almost equal to .950 in all conditions. To accurately interpret the values of the coverage, a 95% Agresti-Coull confidence interval was derived (Agresti & Coull, 1998). The interval was equal to [.946; .954]. In some conditions, coverage was just below the Agresti-Coull interval, but this was not interpreted as deviant from the rule of thumb. Only for $N = 50$, coverage was substantively lower than we expected.

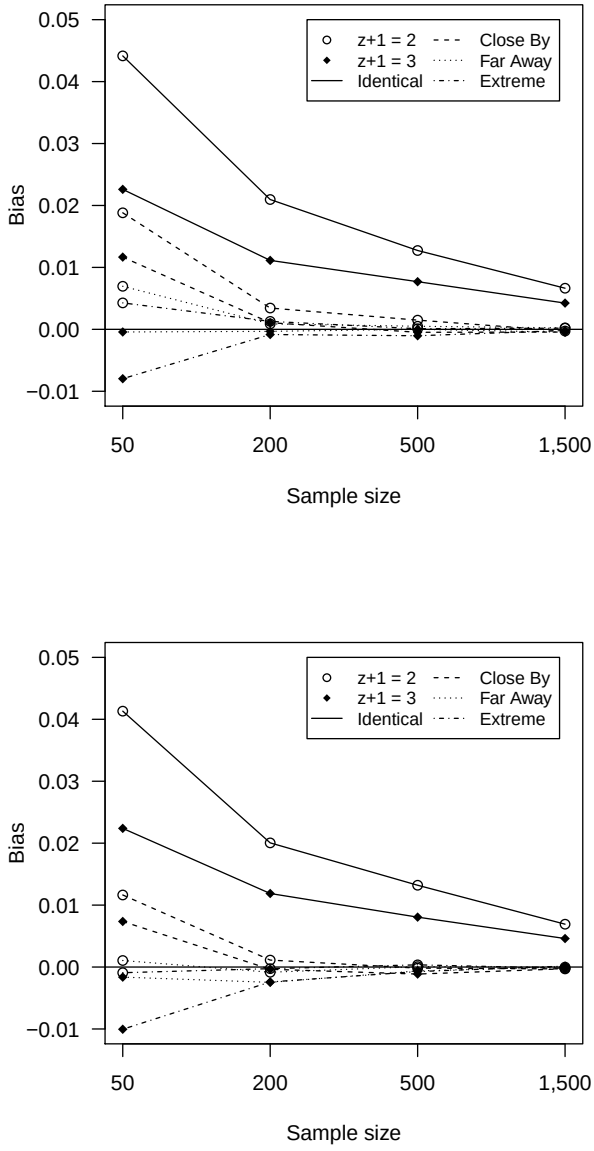

Figure 4.1: *Bias in Scalability Coefficient H for* $J = 2$ *(Upper Panel) and* $J = 3$ *(Lower Panel) for All Four Item Step Orderings.*

Table 4.6: *Results of Simulation Study 1: For Item Step Ordering Identical, Bias of Standard Errors of H (bias.se), and Coverage of 95% CI. Values Outside the Agresti-Coull Interval [.946; .954] Are Printed in Boldface.*

$z+1$	J	N	*bias.se*	95%Cov.
2	2	50	.000	**.929**
		200	.002	.947
		500	.002	.948
		1,500	.001	.952
	3	50	-.002	**.927**
		200	.001	**.943**
		500	.001	**.939**
		1,500	.001	**.943**
3	2	50	-.003	**.924**
		200	.000	**.945**
		500	.000	**.943**
		1,500	.000	.947
	3	50	-.001	**.932**
		200	.000	**.941**
		500	.000	**.941**
		1,500	.000	**.942**

4.4 Simulation Study 2

Simulation study 1 showed that bias was unaffected as number of items increased from two to three but these small test lengths were deemed insufficient for ruling out bias effects for larger J. Study 1 also showed that the bias of estimated coefficient H decreased as number of answer categories increased, which was inconsistent with our expectations. Thus, for larger number of items ($J = 10$) and larger number of answer categories ($z + 1 = 5$), we investigated the bias of coefficient H and its standard error, and the coverage of the 95% confidence interval in a second simulation study.

Study 2 was done only for design cells having identical item steps, because the results of Study 1 showed that for the other distances between item steps bias of the estimates of H and the standard errors was negligible, and coverage was close to .950. The design of the second simulation study was similar to

that of Study 1, and data were generated similarly. To keep the second study manageable, we did not fully cross all factors.

For Simulation study 2, the population values for scalability coefficient H were equal to .293 for ten dichotomous items, .327 for ten trichotomous items, and independent of the value of J .369 for items with five answer categories. Because the item step response functions were identical for all items, the population values stay the same as sample size or number of items increases.

4.4.1 Results

Table 4.7 shows the results of Study 2. The bias of the estimates and the bias of the standard errors of scalability coefficient H were unaffected as J increased from three to ten; bias values were comparable to those found in Simulation study 1. When the number of answer categories increased from three to five, the bias of the estimates decreased, especially for $N = 50$. This outcome again contradicts our expectation that bias increases as $z + 1$ increases.

Figure 4.2 shows the coverage for $J = 10$ items. Coverage of the 95% confidence interval was substantially lower for $J = 10$ than for $J = 3$. None of the values lay in the Agresti-Coull interval. These results contradict our expectation that the coverage remains the same as the number of items increases. Coverage of the 95% confidence interval was even worse for dichotomous items than for polytomous items with three or five answer categories, which is inconsistent with our expectation that coverage deteriorates as number of answer categories increases. Consistent with the results from the first study, we found that coverage was considerably lower for $N = 50$ than for larger sample sizes.

4.5 Discussion

The estimates and the standard errors of Mokken's scalability coefficients are based on the assumption that the sample ordering of the item steps is identical to the ordering in the population. A violation of this assumption may bias the estimates and standard errors of scalability coefficients and

Table 4.7: *Results of Simulation Study 2: For Identical Item Steps for* $J = 10$ *Items (Upper Panel) and for* $z+1 = 5$ *Answer Categories (Lower Panel), Bias of Estimates of H (bias), Bias of Standard Errors of H (bias.se), and Coverage of 95% CI. Values Outside Agresti-Coull Interval [.946; .954] Are Printed in Boldface.*

$z+1$	J	N	*bias*	*bias.se*	95%Cov.
2	10	50	.040	-.001	**.901**
		200	.020	.000	**.902**
		500	.013	.000	**.893**
		1,500	.007	.000	**.896**
3	10	50	.023	-.002	**.921**
		200	.012	.000	**.927**
		500	.008	.000	**.929**
		1,500	.005	.000	**.933**
5	2	50	.016	-.003	**.920**
		200	.008	.000	**.939**
		500	.005	.000	**.944**
		1,500	.003	.000	**.944**
5	3	50	.013	-.003	**.928**
		200	.007	.000	**.944**
		500	.005	.000	.946
		1,500	.003	.000	**.944**
5	10	50	.016	-.001	**.931**
		200	.007	.000	**.937**
		500	.005	.000	**.937**
		1,500	.003	.000	**.941**

produce incorrect coverage of the corresponding confidence intervals. In two simulation studies, we investigated the effect of differences between sample and population item step orderings on the magnitude of the bias and on the coverage of the 95% confidence intervals.

In almost all conditions bias of the estimates of H was negligible, which indicates that the heuristic guidelines of Sijtsma and Molenaar (2002, p. 56) may be too strict. The results suggest that only if item steps are identical or if sample size is very small ($N < 200$), a very small positive bias may

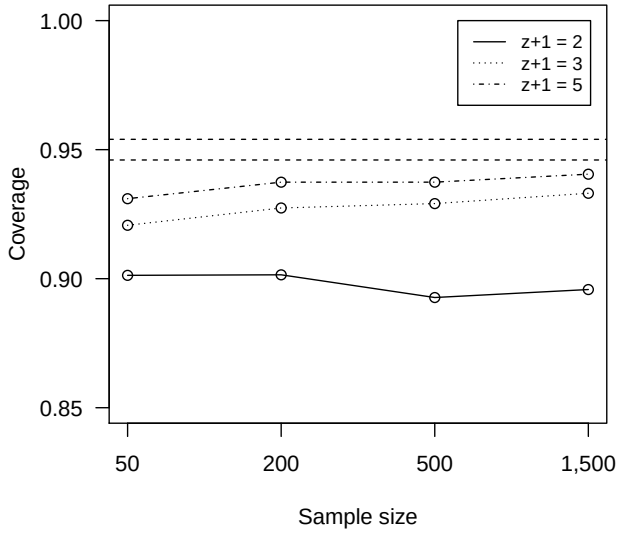

Figure 4.2: *Coverage of 95% Confidence Intervals for $J = 10$ for Varying Number of Answer Categories.*

be expected. Inconsistent with our expectations, bias in the estimates of H decreased as the number of answer categories increased. We have no satisfactory explanation for this phenomenon, but the decrease in bias was persistent when number of answer categories was raised to 5 in Study 2. For all other conditions in Study 2, bias values of the estimates were comparable to those found in Study 1.

Bias of the standard errors of H was negligible in all design cells. Hence, it seems that the categorical marginal modelling approach is an accurate method for deriving standard errors of scalability coefficients. Although the method is rather involved, the implementation in the R-package *mokken* makes the standard errors readily accessible to a general audience.

For most conditions, coverage of the 95% confidence intervals was slightly under .950. For small samples and large numbers of items coverage was slightly poorer. For dichotomous items the coverage dropped to 90% for

large item sets. Although poor coverage in case of large item sets was consistent with our expectations, we did not have an explanation as the bias of the estimates and the bias of the standard errors was unaffected. The coverage may be related to the skewness of the distribution of the sample estimates of coefficient H over replications. For the design cell having the worst coverage, the distribution of the sample estimates of H was positively skewed (skewness equalled .144; computed using the R-package *e1071* by Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2014), whereas the skewness was approximately 0 for design cells that resulted in a correct coverage. In addition, we performed a Kolmogorov-Smirnov test to assess whether the distribution of the estimates of H deviated from normality. All tests were significant, implicating that for large sets of items the distribution of H indeed deviates from a standard normal distribution, which thus may have affected the coverage of the 95% confidence intervals for these conditions. Although even the worst coverages of the 95% confidence intervals may be adequate for practical use, these coverages may be improved if asymmetric confidence intervals are used. The Wald-based 95% confidence interval used in this study (i.e., $\widehat{H}_q \pm 1,96 * se(\widehat{H}_q)$) is symmetric by definition, whereas confidence intervals such as likelihood profile confidence intervals or score confidence intervals (e.g., Lang, 2008) or bootstrap confidence intervals (e.g., Efron & Tibshirani, 1993) can be asymmetric and may improve the coverage. This is a topic for further research.

Other important topics for future research include investigating how the use of standard errors in item selection affects the automatic item selection procedure (Sijtsma & Molenaar, 2002, Chapter 4) in Mokken scale analysis. Now, this procedure only uses the sample values of the scalability coefficients for constructing Mokken scales. However, ignoring standard errors of the scalability coefficients seems to be a source of selection error (Kuijpers et al., 2013b). Taking into account the uncertainty of the estimates in the automated item selection procedure would cause more well-discriminating items to be included in a scale, but this has not yet been systematically investigated. A related topic for future research is the implementation of the standard errors in the automatic item selection procedure in Mokken scale analysis.

Chapter 5

Comparing Estimation Methods for Categorical Marginal Models

Abstract Categorical marginal models are flexible models for modelling dependent or clustered categorical data without making any specific assumptions about the nature of the dependencies. Categorical marginal models are used for different purposes, including hypothesis testing, assessing model fit, and regression problems. Two different estimation methods are used to estimate the marginal models: maximum likelihood (ML) and generalized estimating equations (GEE). We explored three different cases to find out to what extent the two types of estimation methods are appropriate for investigating different types of research questions. The results suggest that ML may be preferred for assessing model fit because GEE has limited fit indices, whereas both methods can be used to assess the effect of independent factors in regression. Moreover, ML is asymptotically efficient, while GEE loses efficiency when the working correlation matrix is not correctly specified. However, for parameter estimation in regression GEE is easier to apply from a computational perspective.

This chapter has been accepted for publication as Kuijpers, R. E., Bergsma, W. P., Van der Ark, L. A., & Croon, M. A. (in press). Comparing estimation methods for categorical marginal models. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research: The 78th Annual Meeting of the Psychometric Society* (pp. 359-376). New York: Springer.

5.1 Introduction

In the social and behavioral sciences, researchers frequently collect data that are correlated or dependent, such as longitudinal data, dyadic data, and data obtained from psychological or educational testing in which each respondent answers several items. Although the dependencies are not always of main interest for the research, they cannot be ignored. Ignoring the dependencies in the analysis may produce incorrect standard errors and p-values. Categorical marginal models (Bergsma et al., 2009) are flexible models for categorical data that take these dependencies into account without making assumptions about their nature. These models are useful when researchers investigate research questions concerning the marginal distributions of a set of variables instead of testing hypotheses with respect to the joint distribution for all variables in a certain data set.

Categorical marginal models are used to answer various types of research questions. Two types of research questions we encountered in the literature are research questions that involve hypothesis testing and research questions that involve parameter estimation. An example of a research question that involves hypothesis testing is provided by Kuijpers, Van der Ark, and Croon (2013a). They proposed fitting categorical marginal models to test the hypothesis that Cronbach's alpha is equal for two or more subgroups. Other examples include testing categorical marginal models for scalability coefficients (Kuijpers et al., 2013b, Van der Ark et al., 2008a), marginal homogeneity (Bergsma et al., 2009), and ordinal association measures (e.g., Lang, 2004).

For the second type of research question, the main interest lies in the values of the estimated regression parameters. For example, Molenberghs and Verbeke (2005) used marginal models to investigate the effect of two types of vaccinations from two different companies on the presence or absence of headaches and respiratory problems in two trial periods. Other examples include 1) modelling the effect of different demographic variables on the relation between smoking and drinking behavior in different subgroups of the Belgian Interuniversity Research on Nutrition and Health study (Kesteloot, Geboers, & Joossens, 1989) and 2) investigating whether different (combinations of) variables such as gender, age, education, and religiosity have a

significant effect on the attitude towards women's roles (Bergsma et al., 2009, pp. 168-171).

Both likelihood methods and quasi-likelihood methods have been used to estimate marginal models. For likelihood methods, which include maximum likelihood (ML) estimation (Bergsma, 1997), maximum empirical likelihood (MEL) estimation, and maximum augmented empirical likelihood (MAEL) estimation (Van der Ark et al., 2011, 2013), the full likelihood is optimized under the marginal model of interest and under the assumption that the data follow a multinomial distribution. ML, MEL, and MAEL estimation differ with respect to whether or not they use all possible item-score patterns of a set of items for the estimation of a model. For research questions that concern hypothesis testing, the authors have used ML (e.g., Kuijpers et al., 2013a, 2013b; Van der Ark et al., 2008a). For this chapter, we only consider ML estimation. The most popular quasi-likelihood method is generalized estimating equations (GEE; Liang & Zeger, 1986). GEE is not based on a specific probability model for the data. The estimation method assumes only a mean-variance relationship for the dependent variable. GEE is mainly used for estimating regression models (e.g., Agresti, 2013; Molenberghs & Verbeke, 2005; Pawitan, 2001). Skrondal and Rabe-Hesketh (2004, p. 200) noted that GEE has some limitations with respect to hypothesis testing and assessing model adequacy.

In this study, we explored to what extent ML estimation and GEE are appropriate for investigating three types of research questions. We considered three different research questions, referred to as Case 1, Case 2, and Case 3. Let θ denote a particular coefficient, and let c denote a fixed value. In this study θ can refer to either the mean (μ) or the reliability coefficient Cronbach's alpha (α). In Case 1, we investigated whether θ is equal to a fixed value c (i.e., $\theta = c$); in Case 2, we investigated whether θ is equal for two groups (i.e., $\theta_1 = \theta_2$); and in Case 3, we investigated whether θ is a linear function of independent variable X (i.e., $\theta = \beta_0 + \beta_1 X$). In each case, we investigated the two coefficients μ and α, and we compared the results obtained with ML estimation and GEE. We illustrated each case with a real-data example.

The remainder of this chapter is organized as follows. First, we briefly

explain categorical marginal models. Second, we discuss the two groups of estimation methods. Third, we discuss how to express μ and α in an appropriate notation for ML estimation. Fourth, using a real-data set, we compare the estimation methods for the three cases. Finally, we discuss the outcomes, and provide recommendations for future research.

5.2 Categorical Marginal Models

In order to use categorical marginal models for testing hypotheses for a coefficient or for estimating parameters in a regression model, the first step is to write the coefficient or the regression model as a function of the frequencies of the item-score patterns that are observed in the data. Consider a set of J items, each item having $z+1$ ordered answer categories $(0, 1, \ldots, z)$; this produces $L = (z+1)^J$ possible item-score patterns. Let $\mathbf{n}$ be an $L \times 1$ vector containing the observed frequencies of the L possible item-score patterns. For example, a dichotomously scored test consisting of $J = 3$ items (denoted by a, b, and c) has $L = 2^3 = 8$ possible item-score patterns; hence vector $\mathbf{n}$ equals

$$\mathbf{n} = \begin{pmatrix} n_{abc}^{000} \\ n_{abc}^{001} \\ n_{abc}^{010} \\ n_{abc}^{011} \\ n_{abc}^{100} \\ n_{abc}^{101} \\ n_{abc}^{110} \\ n_{abc}^{111} \end{pmatrix}, \tag{5.1}$$

where the subscripts denote the items and the superscripts the item scores. The observed frequencies of the item-score patterns in vector $\mathbf{n}$ are given in lexicographic order, running from $00\ldots0$ to $zz\ldots z$ with the last digit changing fastest and the digit in the first column changing slowest.

The expected frequencies under a categorical marginal model are collected in an $L \times 1$ vector $\mathbf{m}$. Because there may be more than one set of expected frequencies that satisfy a marginal model, $\mathbf{m}$ is as close as possible to $\mathbf{n}$. Let matrix $\mathbf{C}$ be a *marginal matrix* consisting of zeros and ones, such that $\mathbf{C}'\mathbf{m}$ produces the relevant marginals from the contingency table. Vector $\boldsymbol{\beta}$ contains the K model parameters β_k $(k = 0, 1, \ldots, K-1)$. Then, let $\mathbf{Z}$ be

the design matrix of the marginal model that uses effect coding in order to select the right parameters from vector $\boldsymbol{\beta}$. In a categorical marginal model, a function of the relevant marginals is then written as

$$\mathbf{f}(\mathbf{C}'\mathbf{m}) = \mathbf{Z}\boldsymbol{\beta}, \tag{5.2}$$

where $\mathbf{f}$ is an appropriate vector function. Alternatively, the model can be written without parameter vector $\boldsymbol{\beta}$ (Agresti, 2013, pp. 460-461; Aitchison & Silvey, 1958; Bergsma, Croon, & Hagenaars, 2013). Let $\mathbf{B}$ be the orthogonal complement of $\mathbf{Z}$, then $\mathbf{B}'\mathbf{Z} = 0$. By premultiplying both sides of Equation 5.2 by $\mathbf{B}'$, the categorical marginal model can be written as a set of constraints

$$\mathbf{B}'\mathbf{f}(\mathbf{C}'\mathbf{m}) = \mathbf{B}'\mathbf{Z}\boldsymbol{\beta} = \mathbf{0}.$$

Because $\mathbf{B}$ and $\mathbf{C}$ are known design matrices, we can write $\mathbf{g}(\mathbf{m}) = \mathbf{B}'\mathbf{f}(\mathbf{C}'\mathbf{m})$. Then, a concise notation for a categorical marginal model, as is used throughout the literature (e.g., Bergsma, 1997; Kuijpers et al., 2013a, Van der Ark et al., 2008a), is

$$\mathbf{g}(\mathbf{m}) = \mathbf{0}. \tag{5.3}$$

Let D be the number of constraints on the expected frequencies $\mathbf{m}$. Each constraint is a scalar function, so for example $g_1(\mathbf{m}) = d_1$, and can be collected in the vector $\mathbf{g}(\mathbf{m})$. So $\mathbf{g}(\mathbf{m})$ contains all constraints that are placed on a vector $\mathbf{m}$. The constraints in Equation 5.3 constitute the categorical marginal model. Some examples of constraints are $\alpha = .80$ and $\mu_1 = \mu_2$.

5.3 Estimation Methods

5.3.1 Likelihood Methods

Likelihood methods use the constraint notation in Equation 5.3 in combination with ML estimation. The unconstrained log-likelihood function (for more details see Bergsma, 1997) is

$$\ell(\mathbf{m}|\mathbf{n}) = \mathbf{n}' \log \mathbf{m}.$$

The maximum likelihood estimate $\widehat{\mathbf{m}}$ maximizes $\ell(\mathbf{m}|\mathbf{n})$ subject to the constraints implied by the categorical marginal model, $\mathbf{g}(\mathbf{m}) = \mathbf{0}$ (Equation 5.3),

and to the constraint that $\sum_i m_i = \sum_i n_i = N$, where N denotes the total sample size.

Let $\boldsymbol{\lambda}$ be a $D \times 1$ vector of Lagrange multipliers and let ν be a single Lagrange multiplier, then under some regularity conditions, the ML estimates under Equation 5.3 are a saddle point of the Lagrangian log-likelihood

$$\ell(\mathbf{m}|\mathbf{n}, \boldsymbol{\lambda}, \nu) = \mathbf{n}' \log \mathbf{m} - \nu(\mathbf{1}'\mathbf{m} - N) - \boldsymbol{\lambda}'\mathbf{g}(\mathbf{m}). \tag{5.4}$$

Bergsma (1997) proposed a Fisher scoring algorithm to find the vector $\mathbf{m}$ in Equation 5.4. The fit of the categorical marginal model can be assessed by means of a likelihood ratio test $G^2 = 2\mathbf{n}' \log(\mathbf{n}/\widehat{\mathbf{m}})$ or a Pearsons' chi-square test $X^2 = (\widehat{\mathbf{m}} - \mathbf{n})'\mathbf{D}_{\widehat{\mathbf{m}}}^{-1}(\widehat{\mathbf{m}} - \mathbf{n})$ with D degrees of freedom. Here, $\mathbf{D}_{\widehat{\mathbf{m}}}$ is a diagonal matrix with the elements of vector $\widehat{\mathbf{m}}$ on the diagonal. Because ML estimation is based on the likelihood function, models can be compared and statistical inferences about parameters can be made.

5.3.2 GEE

GEE specifies a link function for the mean, and specifies the dependence of the variance on the mean. Furthermore, GEE replaces the often complex dependence structure by a so-called *working correlation* structure that is more straightforward to define. GEE can be used to fit any categorical marginal model expressed in terms of Equation 5.2, but traditionally GEE is used for regression models for longitudinal data. In the case of longitudinal data, Y_{it} is the response for person i (with $i = 1, 2, \ldots, N$) on time point t (with $t = 1, 2, \ldots, T$). For GEE, for person i, the model of interest is equal to

$$h(\boldsymbol{\mu}_i) = \mathbf{Z}_i\boldsymbol{\beta}, \tag{5.5}$$

In Equation 5.5, $h(\cdot)$ is a link function that applies element by element to vector $\boldsymbol{\mu}_i$. Vector $\boldsymbol{\mu}_i$ contains the expected responses (i.e., for person i, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT})'$).

GEE links the mean μ to a linear predictor and in addition specifies a variance function that describes how the variance of Y_{it} depends on μ_{it} (Agresti, 2013, p. 462). This model applies to the marginal distribution for each Y_{it}. The estimating equation used in GEE is

$$\sum_{i=1}^{N} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{5.6}$$

where $\mathbf{y}_i$ is a vector with t observed responses (i.e., $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$), and $\mathbf{V}_i$ is an appropriately chosen working correlation matrix. The estimates of the parameters β_i in vector $\boldsymbol{\beta}$ are a solution of Equation 5.6. For an exponential family $\mu_{it} = E(Y_{it})$.

For GEE, the particular working correlation structure needs to be specified for the relation between the t different responses of person i collected in $\mathbf{y}_i$. Different correlation structures can be chosen, depending on the nature of the dependencies between the different responses (Pawitan, 2001, p. 396). Choosing a working correlation structure that approximates the true correlation structure between the dependent responses enhances the efficiency of the parameter estimates (Agresti, 2013, p. 463). Commonly used specifications of the working correlation matrix are: (1) the independence structure, which treats the different responses as independent; thus, no dependency exists; (2) the exchangeable structure, which assumes constant dependency; thus, the correlations between the different responses are assumed to be equal for each observed response; (3) the autoregressive structure, which is often used for measurement over time, and treats the correlations as an exponential function of the time lag; thus, this structure assumes that observations farther apart in time have weaker correlations; and (4) the unstructured structure, which assumes a free specification of the working correlation matrix, implying a separate correlation for each pair of observations (see Agresti, 2013, p. 462, and Pawitan, 2001, pp. 396-397, for more details).

The choice of the working correlation structure determines the GEE estimates of the model parameters and the accompanying standard errors (Agresti, 2013, pp. 462-463). However, even if the working correlation matrix is misspecified, the estimates of the parameters are consistent. In contrast, the estimates of the standard errors of the parameters are not accurate, and need to be adjusted for misspecification of the working correlation matrix by using the so-called sandwich estimator (e.g., Agresti, 2013, p. 467). Liang and Zeger (1986) proposed estimating the GEE parameter estimates and the standard errors by means of a Fisher scoring algorithm.

GEE can also be used for fitting categorical marginal models that are defined by more complex functions than the link function $h(\cdot)$, and by functions that have $\mathbf{n}$ rather than $\mathbf{y}$ as an argument. Here, $\mathbf{f}(\mathbf{C}'\mathbf{n})$ is a function

of the observed responses and $\mathbf{Z}\boldsymbol{\beta} = \mathbf{f}(\mathbf{C}'\mathbf{m})$ is a function of the expected responses, so Equation 5.6 becomes

$$\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{f}(\mathbf{C}'\mathbf{n}) - \mathbf{Z}\boldsymbol{\beta}) = \mathbf{0}. \tag{5.7}$$

A marginal model $\mathbf{Z}\boldsymbol{\beta}$ can represent a wide range of parameters or coefficients, with $\mathbf{f}(\mathbf{C}'\mathbf{n})$ being the corresponding sample value (Bergsma et al., 2013). Equation 5.7 can easily be solved by using

$$\boldsymbol{\beta} = (\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{f}(\mathbf{C}'\mathbf{n}), \tag{5.8}$$

which is equivalent to weighted least squares, with $\mathbf{V}^{-1}$ being a weight matrix. By means of Equation 5.8, estimates for the parameters in $\boldsymbol{\beta}$ can be obtained.

5.4 Expressing Item Means and Cronbach's Alpha in Terms of the Generalized Exp-Log Notation

Maximizing the Lagrangian likelihood in Equation 5.4 requires the matrix of first partial derivatives of $\mathbf{g}(\mathbf{m})$ with respect to $\mathbf{m}$. This matrix, also known as the Jacobian, is usually difficult to obtain. However, if $\mathbf{g}(\mathbf{m})$ is written in the so-called generalized exp-log notation (Bergsma, 1997; Kritzer, 1977) the derivation of the Jacobian is straightforward, and an automated recursive algorithm can be used to compute the Jacobian for a particular categorical marginal model (Bergsma, 1997, p. 68).

5.4.1 Item Means in Exp-Log Notation

For testing hypotheses about the means in vector $\boldsymbol{\mu}$, the coefficient should first be rewritten in the generalized exp-log notation. In this recursive exp-log notation let $\mathbf{A}_1$ and $\mathbf{A}_2$ be appropriate design matrices. Then $\boldsymbol{\mu}$ is equal to

$$\boldsymbol{\mu} = \exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{m})). \tag{5.9}$$

Let $\mathbf{R}$ be a $J \times L$ matrix that contains all L possible item-score patterns. The rows of $\mathbf{R}$ correspond to the J different items. The item-score patterns

in $\mathbf{R}$ are from left to right in lexicographic order, running from $00\ldots0$ to $zz\ldots z$ with the digit in the last row changing fastest and the digit in the first row changing slowest, just as is the case in vectors $\mathbf{m}$ and $\mathbf{n}$. Furthermore, let $\mathbf{u}'_L$ be a $1 \times L$ unit row vector. The $[J+1] \times L$ design matrix $\mathbf{A}_1$ is a concatenation of matrix $\mathbf{R}$ and vector $\mathbf{u}'_L$; that is,

$$\mathbf{A}_1 = \begin{pmatrix} \mathbf{R} \\ \mathbf{u}'_L \end{pmatrix}.$$

For a dichotomously scored test consisting of $J = 3$ items (Equation 5.1) this produces

$$\mathbf{A}_1\mathbf{n} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} n_{abc}^{000} \\ n_{abc}^{001} \\ n_{abc}^{010} \\ n_{abc}^{011} \\ n_{abc}^{100} \\ n_{abc}^{101} \\ n_{abc}^{110} \\ n_{abc}^{111} \end{pmatrix} = \begin{pmatrix} \sum X_a \\ \sum X_b \\ \sum X_c \\ N \end{pmatrix}. \tag{5.10}$$

As the first three elements of the right-hand side of Equation 5.10 show, $\mathbf{Rn}$ produces a vector containing the sum of the scores on items a, b, and c across respondents, and $\mathbf{u}'_L\mathbf{n}$ produces the sample size N.

Let $\mathbf{I}_J$ be an identity matrix of order J. Then, the $J \times [J+1]$ design matrix $\mathbf{A}_2$ is a concatenation of matrix $\mathbf{I}_J$ and unit vector $-\mathbf{u}_J$

$$\mathbf{A}_2 = \begin{pmatrix} \mathbf{I}_J & -\mathbf{u}_J \end{pmatrix}.$$

For the three items a, b, and c, substituting the right-hand side of Equation 5.10 for $\mathbf{A}_1\mathbf{n}$, $\exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{n}))$ yields

$$\exp\left[\begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix} \log\begin{pmatrix} \sum X_a \\ \sum X_b \\ \sum X_c \\ N \end{pmatrix}\right] = \begin{pmatrix} \overline{X}_a \\ \overline{X}_b \\ \overline{X}_c \end{pmatrix}. \tag{5.11}$$

Equation 5.11 shows that $\exp(\mathbf{A}_2 \log(\mathbf{A}_1\mathbf{n}))$ produces the mean score for each item in a data set.

5.4.2 Coefficient α in Exp-Log Notation

Kuijpers et al. (2013a) used categorical marginal models for testing different hypotheses about Cronbach's alpha. They showed that Cronbach's alpha, denoted by α, can be written as a function of $\mathbf{m}$ in the generalized exp-log notation:

$$\alpha = \mathbf{A}_5 \exp(\mathbf{A}_4 \log(\mathbf{A}_3 \exp(\mathbf{A}_2 \log(\mathbf{A}_1 \mathbf{m})))), \tag{5.12}$$

where matrices $\mathbf{A}_1$ to $\mathbf{A}_5$ are appropriate design matrices. For the exact specification of the design matrices and more details about the procedure, see Kuijpers et al. (2013a).

5.5 Three Cases

5.5.1 Data

The use of the two different estimation methods to test three different cases is illustrated by means of a data set obtained by administering a questionnaire to $N = 496$ Dutch union members (Van der Veen, 1992). The questionnaire measures the attitudes and opinions on general militancy, and consists of four subscales - General Attitude, Permissibility, Effectiveness, and Intention - which each contain six items. Each of the six items in a subscale refer to different actions union members can engage in, such as a strike, a protest meeting or a street protest. For the subscales Permissibility and Intention, the answer categories range from 0 to 3, and for the subscales General Attitude and Effectiveness the answer categories range from 0 to 4. Table 5.1 shows the item means, and the values for Cronbach's alpha for each subscale.

Coefficient θ is used to express the different hypotheses. In what follows, θ will be replaced by either the mean (μ) or Cronbach's alpha (α). For ML estimation, we used the R package *cmm* (Bergsma & Van der Ark, 2013), and for GEE, we used the R package *geepack* (Yan, Højsgaard, & Halekoh, 2012).

5.5.2 Case 1: $\theta = c$

First, we tested whether the mean value of General Attitude towards a Strike was significantly greater than 1 (sample value 1.383, Table 5.1). Second, we

Table 5.1: *Item Means and Cronbach's Alpha for each Subscale*

Items	Subscales			
	General attitude	Permissibility	Effectiveness	Intention
Strike	1.383	1.208	1.698	1.151
Work-to-rule	2.278	1.556	1.788	1.536
D. walkout	2.266	1.573	1.702	1.442
C. walkout	2.161	1.546	1.560	1.450
Protest meeting	2.653	2.258	1.835	1.589
Street protest	2.214	1.810	1.625	1.351
Cronbach's alpha	0.744	0.840	0.738	0.877

Note: D. walkout = Demonstrative walkout; C. walkout = Collective walkout

tested whether Cronbach's alpha of the subscale Permissibility was significantly greater than .80 (sample value 0.84, Table 5.1). Nunnally (1978, pp. 245-246) argued that tests used for making decisions about groups should have at least a reliability of .80. The research question is of the form $\theta > c$, and the associated null hypothesis is $\theta = c$.

For investigating $\theta = c$ by means of ML estimation, $\theta = c$ should be written in the constraint notation, $g(\mathbf{m}) = \theta - c = 0$. In the generalized exp-log notation, $g(\mathbf{m}) = \theta - c$ equals

$$g(\mathbf{m}) = \begin{bmatrix} 1 & -c \end{bmatrix} \exp\left(\begin{bmatrix} 1 & 0 \\ 1 & -1 \end{bmatrix} \log\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} \theta\right)\right). \tag{5.13}$$

The categorical marginal model estimates vector $\mathbf{m}$ under the constraint $\theta = c$.

Replacing θ in Equation 5.13 by μ (Equation 5.9) and letting $c = 1$, yields the hypothesis $\mu = 1$. In general, G^2 pertains to a two-sided test. Here, the hypothesis is one-sided, so for a significance level of .05 the value of G^2 at the $2 * .05$ significance level is used. Comparing the observed and expected frequencies allowed us to reject the hypothesis ($G^2 = 77.662$, $df = 1$, $p \leq .000$), and conclude that $\mu > 1$. Replacing θ in Equation 5.13 by α (Equation 5.12), and letting $c = .80$ yields the hypothesis $\alpha = .80$. Comparing the observed and expected frequencies allowed us to reject the hypothesis ($G^2 = 9.489$, $df = 1$, $p = .002$), and conclude that $\alpha > .80$. This example illustrates that likelihood methods can be used to investigate research questions of the type $\theta = c$.

For testing whether $\theta = c$ by means of GEE, $\theta = c$ should be written as $\theta = \mathbf{Z}\boldsymbol{\beta}$. It trivially follows that $\mathbf{Z}$ equals the scalar 1, and $\boldsymbol{\beta} = c$, so $\hat{\theta}$ is trivially fixed to c, and the standard error is zero. The software did not provide goodness of fit statistics. Because $\hat{\theta}$ is fixed to c and no model fit statistics are available, we could not use GEE to meaningfully answer research questions that can be cast into $\theta = c$. This is in accordance with Skrondal and Rabe-Hesketh (2004, p. 200), who stated that GEE has limitations with respect to hypothesis testing and assessing model fit.

5.5.3 Case 2: $\theta_1 = \theta_2$

In this example, we considered whether the population means of the two items General Attitude towards a Demonstrative Walkout and General Attitude towards a Collective Walkout were equal. The sample means for the items were 2.266 and 2.161, respectively (see Table 5.1). Furthermore, we investigated whether the alphas of the two subscales Permissibility and Intention were equal. For the subscale Permissibility $\widehat{\alpha} = 0.840$, for subscale Intention $\widehat{\alpha} = 0.877$ (see Table 5.1). This categorical marginal model can be useful when one wants to compare the alphas of two subscales or tests, or for assessing change in reliability over time. Differences between the reliabilities of two alternate test forms can indicate that the two forms differ in content and measure slightly different traits (Nunnally, 1978, p. 231).

For investigating this model by means of ML estimation, $\theta_1 = \theta_2$ has to be rewritten in the constraint notation, $g(\mathbf{m}) = \theta_1 - \theta_2 = 0$. Because the two coefficients we compared are dependent, vector $\mathbf{n}$ first should be premultiplied by $\mathbf{A}_0$, a marginal matrix (Bergsma et al., 2009, pp. 52-56). Multiplication by matrix $\mathbf{A}_0$ yields the marginal frequencies of the item-score patterns for both sets of items separately. Let L_1 and L_2 be the number of possible item-score patterns for which coefficients θ_1 and θ_2 are computed, respectively. Let $\otimes$ denote the Kronecker product. The general form of the $(L_1 + L_2) \times (L_1 L_2)$ matrix $\mathbf{A}_0$ is

$$\mathbf{A}_0 = \begin{pmatrix} \mathbf{I}_{L_1} \otimes \mathbf{u}'_{L_2} \\ \mathbf{u}'_{L_1} \otimes \mathbf{I}_{L_2} \end{pmatrix}.$$

After premultiplying vector $\mathbf{n}$ by $\mathbf{A}_0$, the two coefficients for the two sets of items are computed using design matrices that are constructed as follows.

Let design matrix $\mathbf{A}_q$, with $q = 1, \ldots, v$, be the particular qth design matrix constructed for the particular coefficient. For testing the equality of two coefficients, design matrices $\mathbf{A}_1$ to $\mathbf{A}_v$ are the direct sum of $\mathbf{A}_q$ and $\mathbf{A}_q$. Since for each design matrix $\mathbf{A}_q$ the procedure is the same, it can be expressed in a general form

$$\mathbf{A}_q^* = \mathbf{A}_q \oplus \mathbf{A}_q = \begin{pmatrix} \mathbf{A}_q & 0 \\ 0 & \mathbf{A}_q \end{pmatrix}.$$

For more details, see Kuijpers et al. (2013a).

In the generalized exp-log notation, $g(\mathbf{m}) = \theta_1 - \theta_2$ equals

$$g(\mathbf{m}) = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}. \tag{5.14}$$

The categorical marginal model estimates vector $\mathbf{m}$ under the constraint $\theta_1 - \theta_2 = 0$. Then, vectors $\mathbf{m}$ and $\mathbf{n}$ are compared by means of G^2 in order to assess whether the two coefficients are equal.

If the coefficient of interest is the mean μ, the population means for the two items are denoted by μ_1 and μ_2, and calculated by using Equation 5.9. For testing Case 2, θ_1 and θ_2 in Equation 5.14 should be replaced by μ_1 and μ_2, respectively. Comparing the observed and expected frequencies allowed us to reject the null hypothesis ($G^2 = 5.429$, $df = 1$, $p = .020$), and conclude that the means are significantly different from each other.

If the coefficient of interest is Cronbach's alpha, the population alphas for the two subscales are denoted by α_1 and α_2, and calculated using Equation 5.12. For testing Case 2, θ_1 and θ_2 in Equation 5.14 should be replaced by α_1 and α_2, respectively. Comparing the observed and expected frequencies allowed us to reject the null hypothesis ($G^2 = 8.939$, $df = 1$, $p = .003$), and conclude that the alphas are not equal.

For GEE estimation, constraint $\theta_1 = \theta_2$ must be cast into Equation 5.2. One possibility is defining a regression model with only an intercept β_0, which can be interpreted as the value of the coefficient under the constraint that $\theta_1 = \theta_2$. Let $\mathbf{Z} = \mathbf{u}_2$, then $\theta_1 = \theta_2$ is equivalent to

$$\mathbf{f}(\mathbf{C}'\mathbf{m}) = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \mathbf{u}_2\beta_0.$$

If the vector of sample estimates of θ_1 and θ_2 is represented by $(\hat{\theta}_1, \hat{\theta}_2)'$, then

the estimating equation (Equation 5.7) reduces to

$$\mathbf{u}_2'\mathbf{V}^{-1}\left(\begin{bmatrix}\hat{\theta}_1\\ \hat{\theta}_2\end{bmatrix} - \mathbf{u}_2\beta_0\right) = \mathbf{0}. \tag{5.15}$$

For an arbitrary correlation matrix $\mathbf{V}$ Equation 5.15 reduces to

$$\begin{bmatrix}\hat{\theta}_1\\ \hat{\theta}_2\end{bmatrix} - \mathbf{u}_2\beta_0 = \mathbf{0},$$

which is minimized for $\hat{\beta}_0 = \frac{\hat{\theta}_1+\hat{\theta}_2}{2}$. So the estimated values for θ_1 and θ_2 are then both equal to the mean of the two values. The hypothesis $\theta_1 - \theta_2 = 0$ can be tested by computing the standard errors by means of the sandwich estimator, computing the confidence interval, and then checking whether 0 is included in the interval.

Using GEE for testing the equality of the means of the two items General Attitude towards a Demonstrative Walkout and General Attitude towards a Collective Walkout, the analysis only estimates a mean value for both values and a standard error, model fit statistics are not available. The estimated mean value for the two means is equal to 2.214, which is obtained independent of the correlation structure. The standard error equals 0.037. To test whether the hypothesis of equal means could be rejected, a 95% Wald confidence interval for the difference between the two means (denoted by $\Delta\mu$) was constructed using $\widehat{\Delta\mu} \pm 1,96 * se(\widehat{\Delta\mu})$. Zero was not included in the interval, so the means are significantly different. GEE was also used for testing the equality of the two alphas of the subscales Permissibility and Intention. The mean value for the two alphas equaled 0.859. The standard error equaled 0.013. A 95% confidence interval for the difference between the two alphas was constructed in a way similar to the computation for the means. Zero was not included in the confidence interval, so the alphas are significantly different.

5.5.4 Case 3: $\theta = \beta_0 + \beta_1 X$

Here, the question was whether the Effectiveness of an action can explain the General Attitude towards that action. We used Effectiveness measured for a Strike (denoted by X_1) and a Work-to-rule (X_2) as the explanatory

variables, and General Attitude measured for a Strike (Y_1) and a Work-to-rule (Y_2) as the outcome variables. Hence, we had $T = 2$ actions and $z + 1 = 5$ levels of the exploratory variable. In longitudinal research, one would consider T time points rather than actions. Estimating a regression model in which Cronbach's alpha is the dependent variable seemed artificial from a substantive point of view. Hence, we only investigated Case 3 for μ. However, there are other situations in which testing the effects of one or more (continuous) variables on the value of a particular coefficient is interesting. For instance, using the log-odds ratio as a measure of association, Bergsma et al. (2013) tested whether the association between two categorical variables remained stable over time.

The regression model is $\mathbf{f}(\mathbf{C}'\mathbf{m}) = \mathbf{Z}\boldsymbol{\beta}$ (Equation 5.2), where $\mathbf{f}(\mathbf{C}'\mathbf{m})$ is the $T(z+1) \times 1$ vector of conditional means:

$$\begin{pmatrix} E(Y_1|X_1 = 0) \\ E(Y_2|X_2 = 0) \\ E(Y_1|X_1 = 1) \\ E(Y_2|X_2 = 1) \\ E(Y_1|X_1 = 2) \\ E(Y_2|X_2 = 2) \\ E(Y_1|X_1 = 3) \\ E(Y_2|X_2 = 3) \\ E(Y_1|X_1 = 4) \\ E(Y_2|X_2 = 4) \end{pmatrix}.$$

Matrix $\mathbf{Z}$ is a $T(z+1) \times 2$ design matrix:

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \\ 1 & 4 \\ 1 & 4 \end{pmatrix}.$$

The first column of matrix $\mathbf{Z}$ is a column of ones, and the second column contains the levels of X_1 and X_2. Vector $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ contains the intercept

and the regression parameter. Vector $\mathbf{m}$ refers to the joint distribution of (X_1, X_2, Y_1, Y_2).

For ML estimation, first $\mathbf{C}'$ and $\mathbf{f}$ should be determined. In our example, pre-multiplying $\mathbf{n}$ by the $(T(z+1)^2 \times L)$ marginal matrix

$$\mathbf{C}' = \begin{pmatrix} \mathbf{I}_{z+1} \otimes \mathbf{u}'_{z+1} \otimes \mathbf{I}_{z+1} \otimes \mathbf{u}'_{z+1} \\ \mathbf{u}'_{z+1} \otimes \mathbf{I}_{z+1} \otimes \mathbf{u}'_{z+1} \otimes \mathbf{I}_{z+1} \end{pmatrix}$$

produces the bivariate marginal frequencies of (X_1, Y_1) and (X_2, Y_2). Function $\mathbf{f}$ consists of two design matrices: $\mathbf{A}_1$ and $\mathbf{A}_2$. Let $\mathbf{r}_{z+1}$ be a $(z+1) \times 1$ vector containing scores $0, 1, \ldots, z$; then $\mathbf{A}_1$ is a $2T(z+1) \times T(z+1)^2$ matrix

$$\mathbf{A}_1 = \mathbf{I}_T \otimes \begin{pmatrix} \mathbf{I}_{z+1} \otimes \mathbf{r}'_{z+1} \\ \mathbf{I}_{z+1} \otimes \mathbf{u}'_{z+1} \end{pmatrix}$$

and

$$\mathbf{A}_2 = \mathbf{I}_T \otimes \begin{pmatrix} \mathbf{I}_{(z+1)} & -\mathbf{I}_{(z+1)\cdot} \end{pmatrix}$$

Hence,

$$\mathbf{f}(\mathbf{C}'\mathbf{m}) = \exp\left(\mathbf{A}_2 \log\left(\mathbf{A}_1\mathbf{C}'\mathbf{m}\right)\right).$$

Second, $\mathbf{B}$, the orthogonal complement of $\mathbf{Z}$, should be determined such that $\mathbf{B}'\mathbf{Z} = \mathbf{0}$. Third, the expected categorical marginal model $\mathbf{B}'\mathbf{f}(\mathbf{C}'\mathbf{m}) = \mathbf{0}$ is estimated, producing estimates for vector $\mathbf{m}$. Using this method for maximizing the likelihood includes the constraints, such that the expected frequencies in vector $\widehat{\mathbf{m}}$ sum to N (Agresti, 2013, p. 460). Fourth, the estimates $\widehat{\mathbf{m}}$ are plugged into model $\mathbf{f}(\mathbf{C}'\mathbf{m}) = \mathbf{Z}\boldsymbol{\beta}$, producing $\mathbf{f}(\mathbf{C}'\widehat{\mathbf{m}})$. Fifth, parameters $\boldsymbol{\beta}$ are obtained by solving

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{f}(\mathbf{C}'\widehat{\mathbf{m}}).$$

Finally, the standard errors of $\hat{\boldsymbol{\beta}}$ are computed using the delta method (for more details, see for instance Bergsma et al., 2009, pp. 71-73), so that each individual parameter in $\boldsymbol{\beta}$ can be tested for significance.

The regression model describes the linear relation between the means that are calculated for each dependent variable given the response to the corresponding independent variable (i.e, the means for Y_1 given the different scores on X_1, and the means for Y_2 given the different scores on X_2). Table 5.2 provides the estimates for the parameters in the regression model.

Table 5.2: *Parameter Estimates using ML Estimation*

Parameter	Estimate	Standard Error
β_0	1.003	0.063
β_1	0.471	0.032

Table 5.3: *Parameter Estimates using GEE Estimation*

Parameter	Estimate	Standard Error
β_0	0.921	0.056
β_1	0.522	0.027

The categorical marginal model also tests whether the regression model that assumes a linear relation between the means fits the data. The results of the analysis showed that the linear regression model does not fit the data, with $G^2 = 173.071$, $df = 8$ and $p < .000$, which implies that the means can not be fitted onto a single straight line; thus, there is not a strictly common linear relation between the conditional means of Y_1 and Y_2 given the scores on X_1 and X_2. However, the regression coefficient is significant, meaning that the scores on X_1 and X_2 have a significant effect on the mean scores of Y_1 and Y_2.

Also, GEE was used to test whether the items Effectiveness of a Strike and Effectiveness of a Work-to-Rule predicted the mean response to General Attitude towards a Strike and General Attitude towards a Work-to-Rule. Table 5.3 shows the GEE estimates of the parameters in the regression model, as defined by Equation 5.2. The regression coefficient is significantly different from zero, which indicates that the scores on X_1 and X_2 have a significant effect on the mean scores of Y_1 and Y_2. For the regression problems, alternative model fit statistics exist for GEE (e.g., Lipsitz & Fitzmaurice, 2009, pp. 62-64; Molenberghs & Verbeke, 2005, pp. 160-161) but these statistics were unavailable in the R package *geepack*, so the model fit could not be investigated.

5.6 Discussion

For this study, we explored to what extent the two estimation methods are appropriate for investigating and testing three types of research questions. The two estimation methods, ML and GEE, both have advantages and disadvantages. ML estimation is based on the likelihood function, so that model fit statistics can be obtained, models can be compared and inferences about individual parameters can be made. In contrast to ML estimation, GEE does not assume a specific probability model for the data, but only assumes a mean-variance relationship for the response variable, making it impossible to obtain likelihood based model fit statistics. Furthermore, GEE replaces the often complex dependence structure by a simpler working correlation matrix. Therefore, GEE is more straightforward to compute than ML methods. For a large number of items, in contrast to GEE, using ML estimation becomes problematic, since it uses each cell of the contingency table for computation of the estimates (Bergsma et al., 2013; Van der Ark et al., 2013). However, ML estimation is asymptotically efficient (e.g., Agresti, 2013), whereas GEE is not when the working correlation structure is not correctly specified.

By means of the three cases, we showed that ML estimation has to be preferred when one is more interested in testing hypotheses and assessing the fit of the marginal model. Both methods are appropriate when one investigates the effect of the independent factors in regression models. For Case 1, GEE could not be used. This is in line with Skrondal and Rabe-Hesketh (2004, p. 200) who stated that GEE has limitations with respect to hypothesis testing and assessing model adequacy. An alternative to solve some of the limitations would be to estimate the standard error of the saturated model, and then use a Wald-based confidence interval to assess whether the value c is included in the confidence interval (Lipsitz & Fitzmaurice, 2009, p. 55). Furthermore, since standard goodness of fit statistics are unavailable for GEE, Lipsitz and Fitzmaurice (2009, pp. 62-64) suggested some alternative model fit diagnostics. For Case 2, ML was easier to apply than GEE, and for ML model fit statistics could be obtained right away. For Case 3, we found that GEE was easier to apply than ML from a computational perspective.

ML estimation uses all item-score patterns that are possible for a set of items, so all elements in vector $\mathbf{n}$ are used. ML estimation becomes prob-

lematic for large numbers of items (e.g., Agresti, 2013, p. 462) because the number of elements in vector $\mathbf{n}$ and the size of the design matrices increase rapidly (Bergsma et al., 2013; Van der Ark et al., 2013). For instance, for a set of ten items ($J = 10$) each with five answer categories ($z + 1 = 5$), the number of elements in vector $\mathbf{n}$ is equal to $(z + 1)^J = 5^{10} = 9,765,625$. An alternative is using MEL estimation (Owen, 2001). MEL uses only the observed item-score patterns, so that the zero-frequencies in vector $\mathbf{n}$ can be ignored. MEL uses much less memory space than ML estimation, and as a result it also is computationally less complex. Therefore, computation time is much shorter, and MEL can be used for large numbers of variables. However, for large sparse contingency tables the empty set problem and the zero likelihood problem can occur when using MEL estimation (for details, see Van der Ark et al., 2013; also see Bergsma, Croon, & Van der Ark, 2012), which causes MEL to break down. Van der Ark et al. (2013) proposed maximum augmented empirical likelihood (MAEL) estimation as a solution for the problems with MEL. MAEL uses all observed item-score patterns, plus a few well-chosen unobserved item-score patterns, the choice of which depends on different factors; see Van der Ark et al. (2013) for more details.

For marginal models, GEE and the likelihood methods require further research. We only illustrated the use of both estimation methods by means of three simple cases for two different coefficients. Many more cases and situations can be investigated. The research can be extended to more complex models and to other coefficients. Furthermore, the cases also can be investigated for MEL and MAEL estimation, which can be compared to GEE estimation in order to investigate which method yields more efficient estimates.

Chapter 6

Epilogue

The central theme of this dissertation is the application of categorical marginal models to psychometric problems in test construction. The use of categorical marginal models in test construction is new, with the exception of Van der Ark et al. (2008a) who used marginal models to construct hypothesis tests for scalability coefficients computed for small sets of dichotomous items. Categorical marginal models have been used mainly to solve sociological research questions, for example to investigate the effect of different variables such as gender, age, education, and religiosity on the attitude towards women's roles (Bergsma et al., 2009, pp. 168-171), and to investigate the effect of two types of vaccinations on possible headaches and respiratory problems in two trial periods (Molenberghs & Verbeke, 2005).

Categorical marginal models are potentially attractive for test construction. First, the models are attractive because the majority of the psychological tests and questionnaires used in social science research consist of items with discrete item scores. Second, the models are flexible, meaning that they are based on only a few, rather weak assumptions. Third, the models are well suited for discrete, dependent data. Standard statistical techniques were not available for particular psychometric problems, for example, for deriving standard errors for Mokken's (1971) scalability coefficients. In other situations, standard methods were based on restrictive assumptions limiting their applicability. Excellent examples are provided by most existing statistical tests for Cronbach's alpha (Cronbach, 1951). The best-known test for alpha, the Feldt (1965) test, is statistically correct given that strong assumptions

like compound symmetry, multivariate normality, and homogeneity of variance are satisfied but this is rare in real-data problems. A major advantage of categorical marginal models over standard statistical techniques is that the models are flexible, with only the assumption that the item scores follow a multinomial distribution.

Categorical marginal models are hardly used, not in psychometrics, and not in other fields. There may be five possible reasons for the models' lack of popularity. The first reason is that the models may be too difficult for non-experts to apply to various research problems. For each coefficient, hypothesis, or model, a researcher must construct a new set of design matrices. This is not easily done, it takes much effort to construct each matrix in the correct way. A second reason why categorical marginal models are rarely used may be that the available software package *cmm* is not very user-friendly. The commands and analyses may be hard to understand for researchers that just start to work with marginal models. The third reason may be that marginal models for categorical data are rather unknown; more research is available for marginal models for *continuous* data. The fourth reason may be that categorical marginal models cannot be used for all research problems. For example, the way categorical marginal models handle missing data needs further investigation. Missing data is an important problem that should be tackled in order to make the models better applicable to more research situations. Furthermore, categorical marginal models cannot handle hypotheses that are stated in terms of inequality constraints. Incorporation of this type of constraints in categorical marginal models is practically relevant and important, since many hypotheses can only be stated in terms of inequality constraints. Finally, latent variable models are popular but the possibility for categorical marginal models to handle latent variables is limited. The fifth reason pertains to the difference between the use of categorical marginal models in sociological research cases versus the use in psychometrics. In the latter, the number of variables is usually much larger which can cause a problem that is often referred to as the curse of dimensionality. This means that when the number variables increases, the size of the design matrices increases exponentially, which causes the method to break down or the set of items cannot even be analyzed in the first place (e.g., see Bergsma et al., 2012, 2013, and

Van der Ark et al., 2011, 2013). It thus causes memory capacity problems in the software used, and as a consequence the categorical marginal models cannot be estimated. Sociological research problems use fewer variables and therefore do not run into these problems that easily.

In this dissertation, I contributed to the solutions of these problems. My main goal was showing that categorical marginal models are suitable methods for solving various psychometric problems. The chapters in this dissertation contribute to this goal by showing that the models can very well be applied to different research problems in test construction. Also, I have contributed to the improvement of the software. In Chapter 3, the standard errors for scalability coefficients I derived by means of categorical marginal models are implemented in the R-package *mokken*. Although this was done quite recently, the method is already used by applied researchers (e.g., Adler-Milstein, Everson, & Lee, 2014; Watson et al., 2014). Furthermore, for testing hypotheses about Cronbach's alpha by means of marginal models, I constructed an extensive help-page in the *cmm*-package so as to illustrate the method's use step by step. Finally, I have contributed to controlling for the curse of dimensionality. For the derivation of standard errors of scalability coefficients (Chapter 3), the design matrices could be adjusted such that they were unaffected by the curse of dimensionality. Furthermore, in Chapter 5 I compared two estimation methods for categorical marginal models, which are the standard estimation procedure used in categorical marginal models (maximum likelihood estimation) and another estimation method that is used mainly for continuous models but that can also be applied to discrete data problems (GEE; Liang & Zeger, 1986). Since the standard estimation method can have problems in case of large item sets, the research in Chapter 5 may contribute to finding alternative estimation methods. The majority of the challenges for categorical marginal models remain yet to be solved. Next, I provide suggestions for further research.

A way to make categorical marginal models more well-known is to make the software more user-friendly. For example, the user manual of the *cmm*-package could be improved and extended, like I already did for testing Cronbach's alpha. More examples of real-data analyses could be implemented in the package. Furthermore, workshops could be given in which the use and

applications of categorical marginal models are discussed. This might contribute to an improved accessibility of categorical marginal models and might help researchers using the method to solve their research problems.

Another important way to make categorical marginal models more well-known is to generalize the research about categorical marginal models to other psychometric problems. Since the categorical marginal modelling approach is flexible, based on weak assumptions, and as I showed in this dissertation proven to be an accurate method for deriving measures of uncertainty, the method can easily be generalized to other coefficients and other hypotheses. For example, hypothesis tests can be generalized to testing Cronbach's alpha for more than two groups; other reliability coefficients like the greatest lower bound (Bentler & Woodward, 1980; Jackson & Agunwamba, 1977; Woodhouse & Jackson, 1977) and coefficient λ_2 (Guttman, 1945); measures used in person-fit analysis; methods for assessing test validity; item rest-score correlations; and so on. In addition, research is required to use the standard errors in real-data research. An example is the implementation of the standard errors of Mokken's scalability coefficients in the automated item selection procedure. This procedure divides a multidimensional set of items into one or more unidimensional scales. Now, items are included in a Mokken scale if the sample values of the item scalability coefficients are at least equal to lower bound c. However, the examples in Chapter 3 suggested that this might be too liberal, and using the standard errors in item selection may statistically accommodate this problem.

Solving the curse of dimensionality provides the most important challenge for categorical marginal models. A solution would be to use a different estimation method for testing the hypotheses, like pairwise maximum likelihood (Lindsay, 1988) or maximum augmented empirical likelihood estimation (MAEL; Van der Ark et al., 2011). However, more research is needed for assessing whether these alternatives are appropriate, and if not, which other alternatives are available.

References

Adler-Milstein, J., Everson, J. & Lee, S.-Y. D. (2014). Sequencing of EHR adoption among US hospitals and the impact of meaningful use. *Journal of the American Medical Informatics Association, 21*, 984-991.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Agresti, A. (2013). *Categorical data analysis.* Hoboken, NJ: Wiley.

Agresti, A. & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician, 52*, 119-126.

Aitchison, J. & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29*, 813-828.

Alsawalmeh, Y. M. & Feldt, L. S. (1994). A modification of Feldt's test of the equality of two dependent alpha coefficients. *Psychometrika, 59*, 49-57.

Bech, P., Bille, J., Moller, S. B., Hellström, L. C., & Ostergaard, S. D. (2014). Psychometric validation of the Hopkins Symptom Checklist (SCL-90) subscales for depression, anxiety, and interpersonal sensitivity. *Journal of Affective Disorders, 160,* 98-103.

Bedford, A., Watson, R., Henry, J. D., Crawford, J. R., & Deary, I. J. (2011). Mokken scaling analysis of the Personal Disturbance Scale (DSSI/sAD) in large clinical and non-clinical samples. *Personality and Individual Differences, 50,* 38-42.

Bentler, P. A., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika, 45,* 249-267.

Bergsma, W. P. (1997). *Marginal models for categorical data.* Tilburg, The Netherlands: Tilburg University Press.

Bergsma, W. P., Croon, M. A., & Hagenaars, J. A. (2009). *Marginal models: For dependent, clustered, and longitudinal categorical data.* New York: Springer.

Bergsma, W. P., Croon, M. A., & Hagenaars, J. A. (2013). Advancements in marginal modelling for categorical data. *Sociological Methodology, 43,* 1-41.

Bergsma, W. P., Croon, M. A., & Van der Ark, L. A. (2012). The empty set and zero likelihood problems in maximum empirical likelihood estimation. *Electronic Journal of Statistics, 6*, 2356-2361.

Bergsma, W. P. & Rudas, T. (2002). Marginal models for categorical data. *The Annals of Statistics, 30*, 140-159.

Bergsma, W. P. & Van der Ark, L. A. (2013). cmm: An R-package for categorical marginal models (Version 0.7) [computer software]. Retrieved from: http://cran.r-project.org/web/packages/cmm/index.html.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395-480). Reading, MA: Addison-Wesley.

Bonett, D. G. (2003). Sample size requirements for comparing two alpha coefficients. *Applied Psychological Measurement, 27*, 72-74.

Cavalini, P. M. (1992). *It's an ill wind that brings no good. Studies on odour annoyance and the dispersion of odorant concentrations from industries.* Ph.D. thesis, University of Groningen, The Netherlands.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

De Fruyt, F., De Bolle, M., McCrae, R. R., Terracciano, A., & Costa Jr., P. T., (2009). Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures. *Assessment, 16*, 301-311.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

EVS (2011). European Values Study 2008, 4th wave, Integrated Dataset (EVS 2008). GESIS Data Archive, Cologne, Germany, ZA4800 Data File Version 3.0.0 (2012-04-10).

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30*, 357-370.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 363-373.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45*, 99-105.

Forthofer, R. N. & Koch, G. G. (1973). An analysis for compounded functions of categorical data. *Biometrics, 29*, 143-157.

Gow, A. J., Watson, R., Whiteman, M., & Deary, I. J. (2011). A stairway to heaven? Structure of the Religious Involvement Inventory and Spiritual Well-Being Scale. *Journal of Religion and Health, 50*, 5-19.

Grizzle, J. E., Starmer, C. F., & Koch, G. G. (1969). Analysis of categorical data by linear models. *Biometrics, 25*, 489-504.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10,* 255-282.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction, studies in social psychology in World War II*, Vol. 4 (pp. 60-90). Princeton, NJ: Princeton University Press.

Hakstian, A. R. & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika, 41*, 219-231.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika, 61,* 679-693.

Hendriks Vettehen, P. G. J., Hagemann, C. P. M., & Van Snippenburg, L. B. (2004). Political knowledge and media use in the Netherlands. *European Sociological Review, 20*, 415-424.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Jackson, P. H., & Agunwamba, C. C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika, 42,* 567-578.

Jansen, A. J. G., Essink-Bot, M.-L., Duvekot, J. J., & Van Rhenen, D. J. (2007). Psychometric evaluation of health-related quality of life measures in women after different types of delivery. *Journal of Psychosomatic Research, 63*, 275-281.

Kendall, M. G. & Stuart, A. (1969). *The advanced theory of statistics* (Vol. 1, 3rd ed.). London: Charles Griffin.

Kesteloot, H., Geboers, J., & Joossens, J. V. (1989). On the within-population relationship between nutrition and serum lipids, the birnh study. *European Heart Journal, 10*, 196-202.

Kim, S. & Feldt, L. S. (2008). A comparison of tests for equality of two or more independent alpha coefficients. *Journal of Educational Measurement, 45*, 179-193.

Kingston, A., Collerton, J., Davies, K., Bond, J., Robinson, L., & Jagger, C. (2012). Losing the ability in activities of daily living in the oldest old: A hierarchic disability scale from the Newcastle 85+ study. *PloS One, 7(2)* e31665.

Kraemer, H. C. (1981). Extension of Feldt's approach to testing homogeneity of coefficients of reliability. *Psychometrika, 46*, 41-45.

Kritzer, H. M. (1977). Analyzing measures of association derived from contingency tables. *Sociological Methods and Research, 5,* 35-50.

Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013a). Testing hypotheses involving Cronbach's alpha using marginal models. *British Journal of Mathematical and Statistical Psychology, 66,* 503-520.

Kuijpers, R. E., Van der Ark, L. A., & Croon, M. A. (2013b). Standard errors and confidence intervals for scalability coefficients in Mokken scale analysis using marginal models. *Sociological Methodology, 43,* 42-69.

Lang, J. B. (2004). Multinomial-Poisson homogeneous models for contingency tables. *The Annals of Statistics, 32*, 340-383.

Lang, J. B. (2008). Score and profile likelihood confidence intervals for contingency table parameters. *Statistics in Medicine, 27*, 5975-5990.

Lang, J. B. & Agresti, A. (1994). Simultaneously modeling the joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association, 89*, 625-632.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13-22.

Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (2010). Investigating invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*, 578-595.

Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics, 80,* 221-239.

Lipsitz, S. & Fitzmaurice, G. (2009). Generalized estimating equations for longitudinal data analysis. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 43-78). Boca Raton, FL: Chapman & Hall/CRC.

Loner, E. (2008). The importance of having a different opinion. Europeans and GM foods. *European Journal of Sociology, 49*, 31-63.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Maydeu-Olivares, A., Coffman, D. L., Garcia-Forero, C., & Gallardo-Pujol, D. (2010). Hypothesis testing for coefficient alpha: An SEM approach. *Behavior Research Methods, 42,* 618-625.

Maydeu-Olivares, A., Coffman, D. L., & Hartmann, W. M. (2007). Asymptotically distribution-free (ADF) interval estimation of coefficient alpha. *Psychological Methods, 12,* 157-176.

McCrae, R. R., Costa Jr., P. T., & Martin, T. A. (2005). The NEO-PI-3: A more readable revised NEO Personality Inventory. *Journal of Personality Assessment, 84*, 261-270.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2014). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3. [computer software]. Retrieved from http://CRAN.R-project.org/package=e1071

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* The Hague/Berlin: Mouton/De Gruijter.

Mokken, R. J. & Lewis, C. (1982). A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement, 6,* 417-430.

Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden, 12* (37), 97-117.

Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data.* New York: Springer.

Murray, A. L. & McKenzie, K. (2013). Estimating the severity of intellectual disability in adults: A Mokken scaling analysis of the Learning Disability Screening Questionnaire. *Psychological Assessment, 25,* 1002-1006.

Muthén, L. K. & Muthén, B. O. (2010). *Mplus Version 6.1.* Los Angeles, CA: Muthén & Muthén.

Németh, R., & Rudas, T. (2013). On the application of discrete marginal graphical models. *Sociological Methodology, 43(1),* 70-100.

Nitschke, J., Osterheider, M., & Mokros, A. (2009). A cumulative scale of severe sexual sadism. *Sexual Abuse: A Journal of Research and Treatment, 21,* 262-278.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Ommundsen, R., Mörch, S., Hak, T., Larsen, K. S., & Van der Veer, K. (2002). Attitudes toward illegal immigration: A cross-national methodological comparison. *The Journal of Psychology, 136*, 103-110.

Owen, A. B. (2001). *Empirical likelihood.* London: Chapman & Hall/CRC.

Pawitan, Y. (2001). *In all likelihood: Statistical modelling and inference using likelihood.* Oxford: Clarendon Press.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika, 31*, 9-12.

R Core Team (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *Journal of Statistical Software, 48(2)*, 1-36.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph, No. 17.*

Samejima, F. (1972). A general model for free-response data. *Psychometrika Monograph, No. 18.*

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350-353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74(1),* 107-120.

Sijtsma, K., & Emons, W. H. M. (2011). Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research, 70(6),* 565-572.

Sijtsma, K. & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Pearson Education.

Van Abswoude, A. A. H., Van der Ark, L. A., & Sijtsma, K. (2004). A comparative study of test data dimensionality assessment procedures under nonparametric IRT models. *Applied Psychological Measurement, 28*, 3-24.

Van der Ark, L. A. (2001). Relationships and properties of polytomous item response theory models. *Applied Psychological Measurement, 25*, 273-282.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software, 20*, 1-19.

Van der Ark, L. A. (2012). New developments in Mokken scale analysis. *Journal of Statistical Software, 48,* 1-27.

Van der Ark, L. A. & Bergsma, W. P. (2010). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75*, 272-279.

Van der Ark, L. A., Bergsma, W. P., & Croon, M. A. (2013). Augmented empirical likelihood estimation of categorical marginal models for large sparse contingency tables. *Under review.*

Van der Ark, L. A., Croon, M. A., & Bergsma, W. P. (2011). *Categorical marginal models for large data sets.* Paper presented International Meeting of the Psychometric Society, Hong Kong, July 2011.

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008a). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika, 73*, 183-208.

Van der Ark, L. A., Croon, M. A., & Sijtsma, K. (2008b). Possibilities and challenges in Mokken scale analysis using marginal models. In K. Shigemasu, A. Okada, T. Imaizuma, & T. Hodhina (Eds.), *New trends in psychometrics* (pp. 525-532). Tokyo: Universal Academic Press.

Van der Veen, G. (1992). *Principes in praktijk: CNV-leden over collectieve acties* [Principles into practice. Labour union members on means of political pressure]. Kampen: J.H. Kok.

Van Onna, M. J. H. (2004). Estimates of the sampling distribution of scalability coefficient H. *Applied Psychological Measurement, 28*, 427-449.

Van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika, 65*, 271-280.

Watson, R., Wang, W., Thompson, D. R., & Meijer, R. R. (2014). Investigating invariant item ordering in the Mental Health Inventory: An illustration of the use of different methods. *Personality and Individual Differences, 66,* 74-78.

Webber, M. P. & Huxley, P. J. (2007). Measuring access to social capital: The validity and reliability of the Resource Generator-UK and its association with common mental disorder. *Social Science and Medicine, 65*, 481-492.

Weijmar Schultz, W. C. M. & Van der Wiel, H. B. M. (1991). *Sexual functioning after gynaecological cancer treatment.* Groningen, The Netherlands: Dijkhuizen Van Zanten B. V.

Weisstein, E. W. (2011). Euler's homogeneous function theorem. From: *MathWorld* – A Wolfram Web Resource http://mathworld.wolfram.com /EulersHomogeneousFunctionTheorem.html.

Woodhouse, B., & Jackson, P. H. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika, 42,* 579-591.

Woodruff, D. J. & Feldt, L. S. (1986). Tests for equality of several alpha coefficients when their sample estimates are dependent. *Psychometrika, 51*, 393-413.

Yan, J., Højsgaard, S., & Halekoh, U. (2012). geepack: Generalized Estimating Equation package. (Version 1.1-6) [computer software]. Retrieved from http://cran.r-project.org/web/packages/geepack/index.html.

Yuan, K-H., Guarnaccia, C. A., & Hayslip Jr., B. (2003). A study of the distribution of sample coefficient alpha with the Hopkins Symptom Checklist: Bootstrap versus asymptotics. *Educational and Psychological Measurement, 63,* 5-23.

Zijlstra, W. P., Van der Ark, L. A., & Sijtsma, K. (2011). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics, 36,* 186-212.

Summary

Social scientists use tests and questionnaires to measure different constructs that cannot be observed directly, such as depression, anxiety, intelligence, work satisfaction, or attitudes towards euthanasia or abortion. Frequently, researchers administering tests to respondents assume that the test-takers do not influence each other's responses, thus they assume that the different respondents' answers or observations are independent. However, answers can also be dependent; for example, respondents can be assessed at multiple occasions, respondents can have a personal relation with each other (e.g., mother and daughter or husband and wife), or respondents are members of the same subgroup (e.g., children attending the same school). When observations in a sample are dependent, standard statistical procedures are not appropriate to use and produce biased results. Methods for analyzing dependent data are available, but many of these methods are based on additional assumptions that may not be satisfied in real data so that these methods can only be applied to a limited number of research questions. A solution is to use marginal models for categorical data, which are flexible models that have only a few assumptions.

In this dissertation, categorical marginal models are applied to various research problems in test construction. Standard statistical procedures are often not available, inappropriate to answer the research question at hand, or are based on restrictive assumptions. Categorical marginal models are flexible models for analyzing dependent or clustered categorical data without making specific assumptions about the nature of these dependencies. Categorical marginal models handle dependencies in a data set by analyzing entire item-score patterns rather than analyzing individual item scores. In order to test various hypotheses or models, categorical marginal models impose

restrictions on particular marginals or cells of the contingency table in which the data are collected. The marginal model is tested under the restrictions that are imposed; the expected frequencies are estimated under the restrictions of the marginal model, such that they are as close as possible to the observed frequencies in the sample. Then, the global fit of the marginal model can be assessed using, for example, a likelihood ratio statistic or Pearson's chi-square statistic.

Categorical marginal models can be applied in a wide range of research areas, yet the use of categorical marginal models in test construction is new. In this dissertation, categorical marginal models are used to solve the following psychometric problems in test construction: constructing hypothesis tests for reliability coefficient Cronbach's alpha (Chapter 2), and deriving standard errors for scalability coefficients in Mokken scale analysis (Chapter 3). The bias of the estimates and the bias of the standard errors of these scalability coefficients is investigated in Chapter 4. In Chapter 5, we explored to what extent two types of estimation methods for categorical marginal models are appropriate for investigating different types of research questions that prevail in test construction.

The most frequently used reliability estimation method is Cronbach's alpha; for almost every published psychological test this coefficient is reported. Most researchers only report the point estimate of Cronbach's alpha. Hence, the uncertainty of the estimate is not taken into account, which can lead to incorrect conclusions with respect to the use of the assessed test. In Chapter 2, the categorical marginal modelling approach is used to derive statistical tests for three relevant hypotheses involving Cronbach's alpha: one where alpha equals a particular criterion; a second testing the equality of two alpha coefficients for independent samples; and a third testing the equality of two alpha coefficients for dependent samples. For each of these hypotheses, various statistical tests have been proposed previously. Over the years, new tests have depended on progressively fewer assumptions. The approach we propose for testing the three hypotheses relies on even fewer assumptions, is especially suited for discrete item scores, and can be applied easily to tests containing large numbers of items. Simulation studies, in which the categorical marginal modelling approach was compared to several existing tests,

showed that the tests based on categorical marginal models had the most accurate Type I error rates, particularly in case of dependent samples.

In Chapter 3, standard errors were derived for scalability coefficients that are used in Mokken scale analysis. Mokken scale analysis, among other model assessment methods, includes an item selection algorithm that can be used to partition a set of items into one or more unidimensional scales, possibly leaving one or more items unscalable. Three types of scalability coefficients are used to determine whether or not items form a scale, and as diagnostics to assess the strength of the scales: (1) a coefficient for each pair of items, (2) a coefficient for each item, and (3) a coefficient for the entire scale. As for Cronbach's alpha, scalability coefficients are usually reported without standard errors or other measures of uncertainty. However, ignoring standard errors can lead to incorrect inferences about which items to include in a Mokken scale and the strength of scales. Although some researchers were able to derive standard errors for one of the scalability coefficients, none were able to derive them for all three coefficients; they were only able to derive standard errors for small sets of dichotomous items. By means of categorical marginal models standard errors are derived for all three types of scalability coefficients, for all types of items used in tests and questionnaires including polytomous items, and for large sets of items. In addition, it was demonstrated by means of two real-data examples that ignoring standard errors of scalability coefficients may result in incorrect inferences with respect to the constructed scales.

The estimates and the standard errors of the scalability coefficients are derived assuming that the ordering of the item steps in the sample is identical to the ordering of the item steps in the population. If this assumption is violated, the estimates and the standard errors may be biased. In Chapter 4, by means of two simulation studies the magnitude of the bias of the estimated scalability coefficients and their standard errors was investigated, as well as the coverage of the corresponding 95% confidence intervals. Bias for the standard errors was negligible in all cases. Bias of the estimates was positive but small for identical items steps, and bias decreased when the distance between item steps increased. Furthermore, bias of the estimates decreased as the number of answer categories and sample size increased. Coverage of

the 95% confidence intervals was close to .950 for all cases. Coverage was poor only for small samples and large numbers of items, the latter particularly when items were dichotomous.

For estimating categorical marginal models, different estimation methods can be used. In Chapter 5, we focus on two of the most frequently used methods, which are maximum likelihood and generalized estimating equations (GEE). Both methods have advantages and disadvantages. In contrast to maximum likelihood, which maximizes the likelihood function under the restrictions of the marginal model, GEE does not assume a specific probability model for the data. Therefore, GEE is simpler and computationally more straightforward than likelihood estimation. However, GEE has problems with respect to efficiency and accuracy when estimating standard errors of parameters or coefficients. In Chapter 5, the two estimation methods for categorical marginal models were compared with respect to the appropriateness of investigating different research questions. This was done by means of three cases. It was concluded that the maximum likelihood method can be used for all types of research questions but that the method becomes problematic for large numbers of items. The GEE method is preferred for conventional regression problems but because the method does not readily provide global goodness-of-fit statistics, it is less useful for the type of hypothesis testing as discussed in for instance Chapter 2.

Nederlandse Samenvatting

Onderzoekers in de sociale wetenschappen gebruiken tests en vragenlijsten om verschillende soorten constructen te meten, zoals depressie, angst, intelligentie, werktevredenheid, of iemands houding ten opzichte van euthanasie. Vaak gaan onderzoekers er vanuit dat respondenten elkaar op geen enkele manier beïnvloeden of dat hun antwoorden op geen enkele wijze met elkaar samenhangen, dus er wordt aangenomen dat de observaties of antwoorden onafhankelijk zijn. Echter, het kan ook voorkomen dat de antwoorden afhankelijk zijn, bijvoorbeeld als de respondent niet aan één, maar aan meerdere meetmomenten heeft deelgenomen, of als respondenten lid zijn van dezelfde subgroep (zoals kinderen die naar dezelfde school gaan), of een persoonlijke relatie met elkaar hebben (zoals een echtpaar of een moeder en een dochter die aan hetzelfde onderzoek meedoen). Als observaties afhankelijk zijn, volstaat het niet om de data met standaard statistische methoden te analyseren, aangezien dit vertekende resultaten op kan leveren. Een oplossing is om afhankelijke data te analyseren met behulp van marginale modellen voor categorische data, aangezien deze modellen flexibel zijn weinig assumpties hebben.

In dit proefschrift worden categorische marginale modellen toegepast op verschillende onderzoeksvragen uit de testconstructie. Andere statistische methoden zijn vaak niet toereikend of zijn gebaseerd op teveel assumpties. Categorische marginale modellen zijn flexibele modellen voor het analyseren van afhankelijke of geclusterde data, zonder dat er specifieke assumpties hoeven worden gemaakt over de aard van deze afhankelijkheden. Categorische marginale modellen houden rekening met deze afhankelijkheden door de gehele antwoordpatronen te analyseren in plaats van de afzonderlijke itemscores. Om verschillende modellen en hypothesen te toetsen, leggen

categorische marginale modellen restricties op bepaalde marginalen van de kruistabel waarin de categorische data verzameld zijn. Het marginale model wordt dan getoetst onder de geformuleerde restricties; de verwachte frequenties worden geschat onder de restricties van het marginale model, en wel zo dat de verwachte frequenties zo dicht mogelijk bij de geobserveerde celfrequenties liggen. Daarna kan de globale fit van het marginale model beoordeeld worden aan de hand van een likelihood ratio test, waarin de afstand tussen de verwachte en geobserveerde frequenties wordt beoordeeld.

Categorische marginale modellen kunnen worden gebruikt om een verscheidenheid aan modellen of hypothesen te toetsen. Over het algemeen worden deze modellen gebruikt om inhoudelijke vraagstukken op te lossen; het gebruik van categorische marginale modellen in testconstructie is nieuw. De diverse onderzoekssituaties waarin categorische marginale modellen zijn toegepast op verschillende psychometrische problemen in testconstructie zijn in dit proefschrift: het construeren van verschillende hypothesetoetsen voor betrouwbaarheidscoëfficiënt Cronbach's alfa (hoofdstuk 2), en het afleiden van standaardfouten voor de drie schaalbaarheidscoëfficiënten in Mokkenschaalanalyse (hoofdstuk 3). Voor deze laatste toepassing is in hoofdstuk 4 onderzocht in welke mate de schattingen en de standaardfouten van de schaalbaarheidscoëfficiënten onzuiver zijn. Ook is in dit proefschrift het gebruik van twee verschillende typen schattingsmethoden voor categorische marginale modellen onderzocht voor verschillende typen onderzoeksvragen (hoofdstuk 5).

Cronbach's alfa is de meest gebruikte coëfficiënt voor het schatten van de betrouwbaarheid van een test. In de meeste gevallen rapporteren onderzoekers alleen de puntschatter als ze de betrouwbaarheid van een test willen weergeven. Een nadeel hiervan is dat de onzekerheid van de schatting niet in ogenschouw wordt genomen, wat kan leiden tot verkeerde conclusies met betrekking tot het gebruik van de beoordeelde test. In hoofdstuk 2 worden met behulp van categorische marginale modellen statistische toetsen geconstrueerd voor drie verschillende hypothesen voor Cronbach's alpha: (1) alfa is gelijk aan een bepaald criterium, (2) de alfa's van twee onafhankelijke steekproeven zijn aan elkaar gelijk, en (3) de alfa's van twee afhankelijke steekproeven zijn aan elkaar gelijk. Voor elk van deze hypothe-

sen zijn verschillende statistische toetsen geïntroduceerd. De methode die wij voorstellen heeft minder assumpties dan bestaande toetsen voor Cronbach's alfa, is geschikt voor discrete data, en kan makkelijk worden toegepast op grote aantallen items. Uit simulatiestudies, waarin de methode van categorische marginale modellen werd vergeleken met verschillende bestaande toetsen, bleek dat de methode van categorische marginale modellen veel nauwkeurigere Type I fouten geeft, met name als steekproeven afhankelijk zijn.

In hoofdstuk 3 worden standaardfouten afgeleid voor schaalbaarheidscoëfficiënten die gebruikt worden in Mokken-schaalanalyse. Mokken-schaalanalyse bestaat uit verschillende methoden om Mokken's model van monotone homogeniteit te onderzoeken, waaronder een itemselectiealgoritme dat gebruikt kan worden om een set items op te delen in één of meerdere schalen, waarbij elke schaal één bepaald construct meet. Drie typen schaalbaarheidscoëfficiënten worden in dit algoritme gebruikt om te bepalen of een item in een schaal wordt opgenomen en zo ja, in welke schaal, en om de sterkte van de schaal te beoordelen: (1) een coëfficiënt voor ieder itempaar, (2) een coëfficiënt voor ieder individueel item, en (3) een coëfficiënt voor de gehele schaal. Net als bij Cronbach's alfa, rapporteren toegepaste onderzoekers vaak alleen de puntschatters van deze coëfficiënten, en nemen ze de onzekerheid van de schatting niet in ogenschouw. Hierdoor kunnen zwakke items in een schaal worden opgenomen terwijl ze er niet in thuishoren, en kunnen goed-discriminerende items ten onrechte uit een schaal worden gelaten. In voorgaand onderzoek is het sommige onderzoekers gelukt standaardfouten af te leiden voor één van de schaalbaarheidscoëfficiënten, maar nooit voor alle drie, en alleen voor kleine aantallen dichotome items. Met behulp van categorische marginale modellen werden in hoofdstuk 3 de standaardfouten afgeleid voor alle drie de schaalbaarheidscoëfficiënten, voor polytome items en grote aantallen items. Verder wordt in dit hoofdstuk aangetoond dat het negeren van de standaardfouten kan leiden tot verkeerde conclusies over geconstrueerde schalen.

Bij de berekening van de schaalbaarheidscoëfficiënten wordt aangenomen dat de ordening van de itemstappen in de steekproef identiek is aan de ordening van de itemstappen in de populatie. Echter, als dit niet het geval is,

zouden de schattingen en de standaardfouten van de coëfficiënten onzuiver kunnen zijn. In hoofdstuk 4 werd de mate waarin de schattingen en de standaardfouten van de schaalbaarheidscoëfficiënten onzuiver zijn onderzocht met behulp van twee simulatiestudies. Uit simulaties bleek dat de standaardfouten in alle gevallen zuiver zijn. De puntschatter is enigszins onzuiver als verscheidene itemstappen identiek aan elkaar zijn, naarmate de itemstappen verder uit elkaar liggen worden de schattingen zuiverder. Ook is de puntschatter zuiverder naarmate de steekproefgrootte toeneemt. Het bereik van het 95% betrouwbaarheidsinterval ligt dicht bij .950 voor vrijwel alle condities. Het bereik van het 95% betrouwbaarheidsinterval wordt slechter als het aantal items toeneemt; dit is vooral zo voor dichotome items.

Voor het schatten van categorische marginale modellen zijn verschillende schattingsmethoden beschikbaar. In hoofdstuk 5 hebben we ons beperkt tot de twee meest gebruikte methoden: maximum likelihood en generalized estimating equations (GEE). Beide methoden hebben voor- en nadelen. Zo heeft GEE problemen met efficiëntie en nauwkeurigheid bij het schatten van standaardfouten van parameters of coëfficiënten, maar is rekenkundig gezien eenvoudiger en makkelijker uit te voeren dan maximum likelihood. Een groot voordeel van maximum likelihood ten opzichte van GEE is echter dat de eerste methode gebaseerd is op een likelihoodfunctie, en dus zijn er grootheden beschikbaar om de fit van een model te beoordelen, wat niet het geval is bij GEE. Voor verschillende typen onderzoeksvragen werd met behulp van drie casussen onderzocht welke schattingsmethode het meest geschikt en het meest bruikbaar is, en welke methode de meest accurate resultaten geeft. Zo blijkt dat maximum likelihood het beste te gebruiken is in het geval van hypothesetoetsen voor verschillende coëfficiënten en het beoordelen van de fit van modellen, en dat GEE het beste te gebruiken is voor het schatten van parameters en regressiecoëfficiënten in regressiemodellen.

Dankwoord

Na vier jaar is het dan eindelijk zo ver, mijn proefschrift is klaar! Via deze weg wil ik iedereen bedanken die op een of ander manier heeft bijgedragen aan deze prestatie.

Allereerst mijn promotor Klaas Sijtsma en mijn copromotoren Andries van der Ark en Marcel Croon: bedankt voor de prettige samenwerking en jullie vertrouwen in mij. Klaas, bedankt voor je kritische blik, minutieuze feedback tijdens het schrijfproces en je (soms vaderlijke) adviezen. Marcel, bedankt voor al je kennis en de aansturing op het technische vlak, vaak heb ik vol bewondering toegekeken hoe jij moeiteloos een blaadje volschreef met vergelijkingen en formules en zo weer een wiskundig probleem de wereld uit hielp. Andries, bedankt voor je dagelijkse begeleiding, je enthousiasme, je feedback, en het feit dat ik altijd bij je terecht kon. Als ik weer eens ongeduldig was en vond dat het allemaal te lang duurde, of als ik vast zat met een probleem van welke aard dan ook bleef jij rustig, bood je een luisterend oor en dacht je mee aan een oplossing.

Mijn promotiecommissie wil ik bij deze bedanken voor het beoordelen van mijn proefschrift. Daarnaast wil ik Wicher Bergsma bedanken voor de fijne samenwerking in en na Londen. Door je scherpe inzichten en kennis heb ik veel bijgeleerd.

Verder wil ik mijn (ex)collega's bij MTO bedanken. In het bijzonder dank ik Daniël P., Florian, Geert, Marcel A., Margot, Marie-Anne en Robert voor de gezellige pauzes, borrels en leuke tijden op congressen. Liesbeth en Marieke, bedankt voor jullie hulp. Pieter, bedankt voor alle gezellige momenten, goede gesprekken en je luisterend oor. Last but definitely not least: bedankt Judith en Gabriëla voor het zijn van de meest leuke kamergenootjes die een mens maar kan hebben, het was een feestje om met jullie op de kamer

te zitten! Ik ben blij dat we, ondanks dat we alledrie niet meer bij MTO zitten, de gezelligheid nog steeds voortzetten. Judith, met je nuchtere kijk kan jij zelfs het grootste wereldprobleem heerlijk relativeren, en je droge humor zorgt ervoor dat ik elke keer weer in een deuk lig. Gabriëla, zoals jij het al zei in jouw dankwoord, ooit zijn we begonnen met kletsen en zijn we er nooit meer mee opgehouden. Niet alleen in goede tijden hebben we gesprekken, vol met verwijzingen naar Sex and the City en Friends ("The end of an era!", "In London?!?"), maar ook in mindere tijden kan ik rekenen op je steun. Ga vooral zo door met je temperamentvolle en gezellige zelf te zijn!

Daarnaast wil ik alle mensen bedanken die ik via IOPS congressen en bestuursvergaderingen heb leren kennen. Willem Heiser, bedankt voor je tips en adviezen, en dat je samen met Henk Kelderman menig IOPS en IMPS congres hebt opgefleurd met leuke verhalen en anekdotes. Verder wil ik alle IOPS-aio's bedanken voor de gezelligheid tijdens de congressen, waar Joost en Dylan als altijd voorzitters van de feestcommissie zijn.

Vicky, bedankt voor het ontwerpen van de kaft van mijn proefschrift! Maarten Kampert, het is altijd weer gezellig je te zien en je krijgt me aan het nadenken over zaken waar ik nog nooit over na had gedacht. Ook zorgt jouw (onbewuste) motto "Om een kort verhaal lang te maken" ervoor dat we niet snel uitgepraat raken!

Ook buiten werk zijn er maar genoeg mensen die mijn leven de afgelopen jaren een stukje leuker hebben gemaakt. HKGD, ofwel Els, Nikki, Roos en Terri, bedankt voor alle gezellige momenten, borrels en uitjes tijdens en na onze studie, en dat jullie altijd naar mijn toneelstukken zijn komen kijken (Titanium!). Kimberly, door jouw gekke grappen en energieke zelf is geen enkele borrel of uitgaansavond saai! Michelle, wij hebben elkaar geloof ik vaker gezien in het buitenland dan in Nederland. Jouw fenomenale gevoel voor humor maakt elke reis hilarisch, dat weet geen (al dan niet denkbeeldige) rat te verpesten.

Jolein en fellow-Aussie fan Patrick, het is altijd weer leuk jullie te zien. Bedankt voor alle hilarische en verrassende avonden en uitjes, goede gesprekken en gastvrijheid in Casa d'n P. Lisette, jouw harde humor is vaak heerlijk op het randje. Naast lachen is er ook plaats voor serieuze gesprekken, en ben ik blij dat we er ook op lastige momenten voor elkaar kunnen zijn.

Daarnaast wil ik uiteraard ook mijn paranimfen Fely en Dabis bedanken voor alle leuke, gezellige, en bijzondere momenten. Jullie hebben me altijd gesteund, en een hart onder de riem gestoken op de momenten dat ik het nodig had. Fely, samen kunnen we uren kletsen over van alles en niets, horoscopen en carnavalsoutfits, maar ook goede gesprekken hebben over Het Leven en De Liefde. Vol humor en met raad en daad sta je me bij, en weet je me ook een spiegel voor te houden waar nodig. Bedankt voor deze fijne vriendschap! Dabis, al zo'n 25 jaar ben je mijn beste vriendin, en ik hoop dat dit zo blijft tot in het bejaardentehuis. Door jou heb ik geleerd nog meer te gaan voor wat je zelf het liefste wilt en je niks aan te trekken van wat anderen vinden. Jouw open blik en humor zorgen ervoor dat ik iets op een manier ga bekijken zoals ik het nog niet bekeken had. Bedankt voor deze hechte vriendschap!

Dan niet te vergeten mijn familie. Karlijn, ook al zijn we het soms niet met elkaar eens, je bent en blijft natuurlijk altijd mijn zus! Het is fijn dat we onze ervaringen in de wetenschap met elkaar kunnen delen, over bepaalde zaken zo hetzelfde kunnen denken, en kunnen grinniken om dingen die niemand anders snapt. Ik ben blij te zien dat je met Matt het ware geluk hebt gevonden, en hoop dat er nog veel mooie tijden voor jullie komen. Matt, ook al kennen we elkaar nog maar kort, ik weet dat er in de VS altijd iemand zal zijn die ik kan bezoeken! Tot slot de pama's. Mama, jij bent zo lekker geordend, daadkrachtig en nuchter. Ik kan altijd bij je terecht, en kan je altijd bellen voor was- en kookvragen. Papa, jouw vrolijkheid en enthousiasme werkt aanstekelijk. Je lacht altijd om je eigen grapjes, en in je overvolle maar o zo opgeruimde schuurtje is er altijd wel een schroefje, latje of plugje te vinden om van alles en nog wat te repareren. Bedankt dat jullie altijd voor me klaarstaan, en dat jullie zo trots op me zijn.